ANNALS

OF THE

NEW YORK ACADEMY OF SCIENCES

VOLUME XLIV

NEW YORK
PUBLISHED BY THE ACADEMY
1943

CONTENTS OF VOLUME XLIV

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
VOLUME XLIV, ART 1. PAGES 1-104
MAY 29, 1943

BOUNDARY-LAYER PROBLEMS IN THE ATMOSPHERE AND OCEAN*

BY

C.-G. ROSSBY, B. HAURWITZ, BENJAMIN HOLZMAN,
WOODROW C. JACOBS, A. A. KALINSKE, PHILLIP LIGHT,
R. B. MONTGOMERY AND H. U. SVERDRUP

CONTENTS

*This series of papers is the result of a conference on Boundary-layer Problems in the Atmosphere and Ocean held by the Section of Oceanography and Meteorology of The New York Academy of Sciences March 6 and 7, 1942.

Publication made possible through a grant from the income of the Nathaniel Lord Britton Fund.

INTRODUCTION TO THE CONFERENCE AND SOME APPLICATIONS OF BOUNDARY-LAYER THEORY TO THE PHYSICAL GEOGRAPHY OF THE MIDDLE WEST

By C.-G. Rossby

Director, Institute of Meteorology, University of Chicago, Illinois

It is with considerable hesitation that I have accepted the honor and responsibility of serving as chairman for this conference on boundary-layer problems in the atmosphere and in the ocean. During the last few years my own activities in this field have been negligible, and it was not until this winter that circumstances led me and my colleagues in the Institute of Meteorology at the University of Chicago to take an active interest in problems of the type that will be discussed in this conference.

The boundary-layer problems of the atmosphere are essentially identical with those problems in which meteorology impinges on the other earth sciences. During the last 10 or 15 years meteorological research in this country has been conducted mainly in engineering institutions and to a considerable extent as an aid to aeronautics. As a result of this association, we have learned to make use of some of the methods of modern fluid mechanics, but our attention has been so directly concerned with the aeronautical applications of meteorology that we may have lost sight of the importance of meteorology in the interpretation of the changes, movements and transfer processes that take place in the solid surface of the earth or in the surface layers of the oceans. In that respect it is to be hoped that the present conference, which is attended by representatives from several of the earth sciences, may help to re-establish meteorology to its real place in geophysics. All I can hope to do in this brief speech is to indicate through a few concrete examples how meteorology might be of some service to other branches of geophysics. This may be illustrated by a number of problems involving the application of atmospheric boundary-layer mechanics to the physical geography of the Middle West, which are being studied at Chicago.

AIRFLOW OVER A SAND DUNE

In the autumn of 1941 we were asked to undertake a study of the airflow over the sand dunes which characterize the eastern shore of Lake Michigan, and particularly to study the transport of sand by wind.

These dunes are roughly parallel to the shoreline and in the main covered with vegetation, but with occasional saddle-shaped cuts in which the sand is exposed and subject to occasional strong drifts.

A field party* began a detailed study of one of these cuts near Benton Harbor, Michigan, in September, 1941 (FIGURE 1). On the meteorolo-

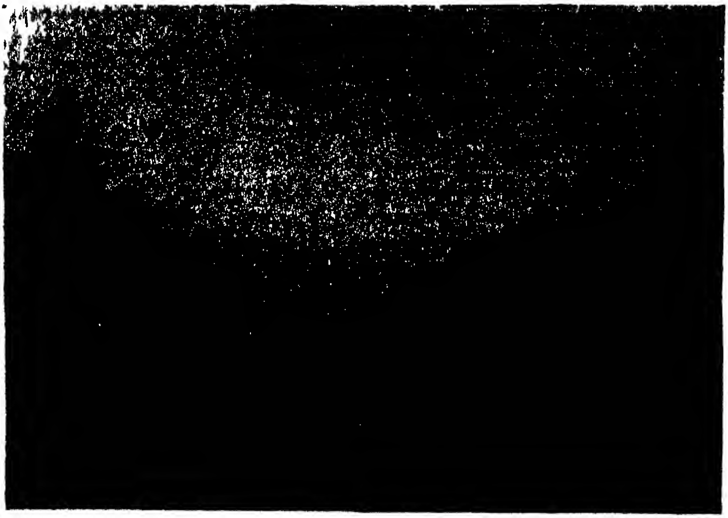


FIGURE 1 View from Lake Michigan of blowout near Benton Harbor

gical side the problems were to determine the general characteristics of the flow pattern over the dune; the characteristic aerodynamic roughness of the surface of the dune, and in general to investigate the stress distribution for different values of the wind at some prescribed reference point; to measure the gustiness of the wind, and the wind-borne transport of sand at different heights above the ground under different wind conditions (different stresses). On the geological side we were interested in the profile of the dune, in the profile, wave length, and speed of the ripple patterns in the surface of the dune, and the relation of these ripples to the wind stress; and in the composition, size and shape of the sand.

Techniques were developed for the measurement of several of these items, but it has been found that a large number of additional measure-

*Directed by Prof R Belknap of the University of Michigan and Prof H Landsberg of the University of Chicago

ments will be needed before any satisfactory analysis of the geological aspects of this problem can be undertaken. A few results of aerodynamic interest can be stated at the present time,* as follows:

1. The flow pattern over the dune and the shape of the dune itself interact in an extremely interesting fashion. Thus, toward the beach the main dune is preceded by a low foredune, behind which a well-developed eddy is observed, with wind motion next to the ground against the prevailing wind. It is not unreasonable to assume that while the foredune now aids in the establishment of this eddy, the foredune itself may be maintained as the result of deposition of suspended sand at the point of stagnation between the reverse eddy flow and the prevailing wind (FIGURE 2).

2. A number of anemometer readings indicate that the small-scale characteristic roughness parameter, z_0, of the sand-dune surface, namely about 0.13 cm., is even less than the roughness of a field of short-cut grass.

3. Over the saddle point of the dune the logarithmic wind-profile does not apply, since the crowding of streamlines at this point produces a maximum wind speed already in a height of about 2 meters above the sand.

4. In other locations, where the logarithmic wind-layer is well developed, it has been possible to determine from large numbers of one-minute wind readings at different levels, and from stress determinations, the correlation (r) between the vertical (w') and down-stream (u') turbulent wind components, i.e., the ratio

$$r = \frac{\overline{u'w'}}{\overline{u'^2}} = -\frac{\tau}{\rho \overline{u'^2}}.$$

In this expression, τ represents stress per unit area and ρ the air density. Values were determined for 7 different heights between 0.5 and 7.7 meters. From about 100 sets of values at each level this correlation was found to have the following reasonably constant values:

Height (cm.)	770	600	500	400	200	100	50
r	0.22	0.23	0.27	0.23	0.28	0.31	0.51

For comparison it might be mentioned that Wattendorf,[4] working with a rectangular channel, obtained a value of about 0.32 from about 3/10 of the distance from the channel center to the wall and up to the immediate vicinity of the wall. Reichert[4] in Göttingen obtained about 0.23. It is of considerable interest that such good agreement is obtained be-

*See H. Landsberg, The structure of the wind over a sand dune Trans Am Geophys, Un. 1942 (2) 237–239.

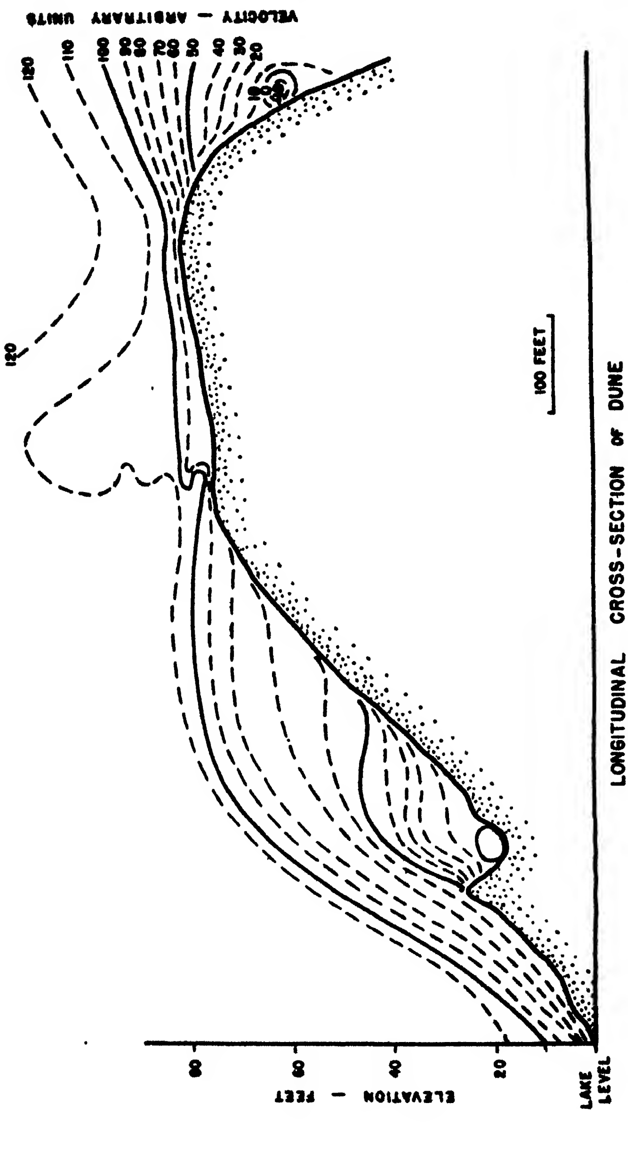


FIGURE 2. Approximate percentage distribution of horizontal wind speed over blowout dune near Benton Harbor. On-shore components positive: opposing winds negative. Notice small eddies behind foredune and on slope of repose.

tween small-scale laboratory experiments and full-scale atmospheric phenomena.

5. The vertical distribution of sand during active drift was measured with the aid of microscope slides exposed at 10-cm. height intervals from 10 cm. up to about 60 cm. It was found that the distribution is in complete disagreement with the concept of turbulent suspension equilibrium. Thus the transport of sand appears to be the result of saltation, turbulent suspension apparently being of negligible consequence, in accord with the results obtained by Bagnold.[1] In one case of fairly strong wind (1415 feet per minute at 5 cm.), the percentage distribution of sand deposited per square cm. was as follows:

Height (cm.)	3	10	20	35	48
Deposit on slide as % of deposit at 10-cm. height	121	100	85	53	22

This shows a nearly linear distribution reaching much greater heights than might be expected on the basis of suspension equilibrium for the fairly coarse sand in the dune.

6. It is our belief that the magnitude of the angle of repose on the leeward of the dune should furnish some information concerning the minimum stress required to set the dune sand in motion.

7. It is of some interest to mention that the total amount of sand transported in a single storm from the beach side to the slope of repose was estimated to be about 90 tons. This drift occurred over a cross-section of 120 square feet and in an interval of about 10 hours, during which the wind remained above the threshold value. This latter quantity was determined for the particular sand involved, which contained about 4 per cent of moisture by weight, to be about 3.5 meters per second at a height of 5 cm. above the sand.

STUDIES OF LAKE MICHIGAN

The second boundary problem of a physical geographical nature concerns the exchange of heat and momentum between Lake Michigan and the overlying atmosphere. There are several reasons why we should be interested in this problem. From a purely scientific standpoint, our interest was aroused by the fact that the Lake Michigan drainage area is small and unimportant and that the lake therefore may be considered as a model for a closed hydrodynamic-thermodynamic system. Practically, the lake is an important factor in Chicago weather. Finally, the lake provides the city of Chicago with a large part of its water supply. The industrial plants along the Indiana shore are now dumping large

amounts of waste products in the lake, resulting in a considerable amount of pollution. It would appear that knowledge of the physical-chemical *characteristics and of the circulation of the water masses in Lake Michigan should have considerable practical value.*

A systematic survey of Lake Michigan water masses is being conducted by Phil Church.[2] Until recently no systematic temperature data from the interior of the lake have been available. With the aid of a recently developed instrument, the bathythermograph,[5,6] which records temperature directly against pressure, it is now a relatively simple matter to collect such data even from moving steamers. A great number of car ferries are operating on fairly regular schedules between Milwaukee or other Wisconsin points on the west side and various Michigan points on the east side. We have been able to operate our bathythermograph equipment from some of these ferries and have made a series of weekly or twice-monthly vertical temperature sections between Milwaukee, Wisconsin and Muskegon, Michigan. The distance between these two ports is about 80 miles and the maximum depth about 60 fathoms. The maximum recorded depth in the lake is, I believe, 154 fathoms, in the latitude of Green Bay but toward the Michigan side. The sections, of which we have about 15 or 16 (two for each round trip, with 20 or 30 soundings in each section), have already revealed some extremely interesting features.

During November and early December there was a marked thermocline separating an upper homogeneous layer from the colder and denser, and nearly homogeneous bottom water (FIGURE 3). In late December, as winter cooling progressed, strong mixing set in, and the thermocline disappeared. Toward the end of the month the water columns were vertically homogeneous. This vertical mixing took place before the water had reached maximum density, suggesting that mechanical, wind-produced stirring played an important role in the cooling process. Warming-up occurred in the bottom layers during this period. Later, the rate of cooling became reasonably uniform throughout the entire depth of the water column. The importance of mechanical stirring is brought out by the fact that by the middle of February the bottom temperature had dropped below the temperature of maximum density. During the late autumn there is a typical horizontal temperature gradient in the section, with two bands of warm water on the eastern and western slopes (the warmest water being found off the Michigan shore) and with a core of distinctly colder water in the middle of the section. In the latter part of the winter, after the lake had cooled below the point of maximum density, the temperature distribution reversed itself, with

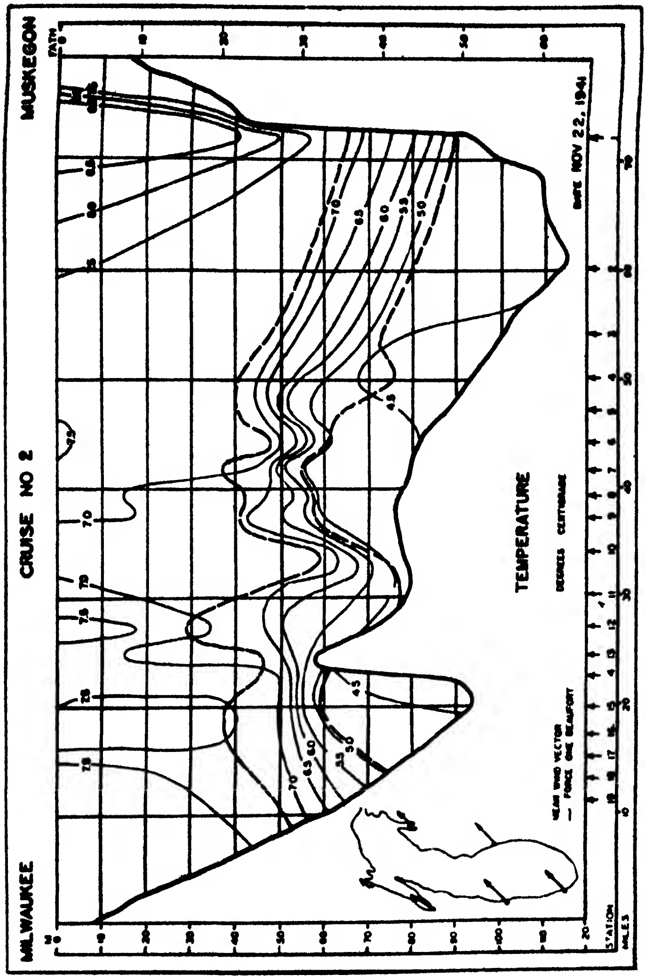


FIGURE 3 Temperature distribution in Lake Michigan in a vertical section between Milwaukee and Muskegon on Nov 22, 1941.

two colder ribbons along the slopes and a warm central core. Thus, there is at all times a tendency toward a horizontal mass distribution with heavy water near the center and lighter water toward the slopes. Assuming that the water masses near the bottom are practically stagnant, this mass distribution suggests a slow cyclonic (counterclockwise) rotation of the upper layers.

About 1890, Harrington[3] organized a study of the surface circulation of Lake Michigan with the aid of drifting bottles. The study was conducted during the summer, whereas our results apply to the cold season. It is nevertheless of interest to point out that Harrington's observations indicate a two-cell type of cyclonic circulation, one occupying the northern and larger portion of the lake and the other the southern *portion, with a boundary somewhere in the vicinity of the line between Milwaukee and Muskegon.*

It is not easy to see how the prevailing southwest or west winds can set up a rotary, cyclonic current system in the lake. It has been suggested that the normally prevailing wind system over Lake Michigan might exert a slight cyclonic torque. During the summer, a quasi-stationary front is likely to be found over the northern part of the lake and during the latter half of November and early December of 1941 we actually did observe a quasi-stationary front extending across the lake, north of Milwaukee, for a considerable part of the period. But the vectorial means of the winds observed at the Weather Bureau stations surrounding the lake do not furnish much support for the hypothesis of a cyclonic wind torque.

In the light of the slow cyclonic circulation suggested by the temperature data, it is difficult to see how the pollution products from the industrial plants are brought to Chicago except through a considerable amount of eddying motion near the tip of the lake, and during the relatively infrequent periods of south or southeast winds. It would seem that a permanent solution to the pollution problem requires much more detailed data on the circulation in the southern part of the lake than are now available.

In the near future we hope to be able to extend our measurements to other parts of the lake and to include also a wider range of physical and chemical data to permit a more accurate description of the lake circulation.

OZONE STUDIES

A third boundary-layer problem in which we are interested concerns the distribution of ozone near the ground. It has been suggested by Wulf[7,8] that during the winter months stratospheric ozone may be brought down into the troposphere in high latitudes and move southward in the continental polar air outbreaks characteristic of the Hudson Bay–Mississippi Valley region. Detection of ozone near the ground would thus furnish proof that Arctic air is formed not only by surface cooling but also by cooling at very high levels. Since ozone is destroyed through

contact with organic material at the ground it therefore must be replenished by turbulent transfer downward. One should therefore expect to find, in Arctic air-masses, a logarithmic vertical distribution of ozone similar to the distribution of wind with height. From the distribution curve it should be possible to determine the rate of diffusion downward and hence the rate of destruction of ozone at the ground. In this fashion one might also gain some information concerning the rate of formation of ozone in the stratosphere. So far we have not been able to overcome the many technical difficulties in the chemical method that has to be used in the ozone measurements needed for this particular problem; hence, it is not possible for me to do more than to point to this problem as a potentially interesting application of boundary layer theory.

SUMMARY

In spite of the very preliminary nature of the comments made above it is my hope that the three problems touched upon might bring out the fact that the laws of fluid mechanics and their application to the structure of the lowest strata of the atmosphere are gradually becoming essential tools not only in the study of the atmosphere itself but also in the study of the impact of the atmosphere on the solid surface of the earth, its vegetative cover and its water masses.

DISCUSSION OF THE PAPER

Dr. C. L. Pekeris (*Columbia University, New York*)

Has any attempt been made to determine the horizontal size of the eddies from the auto-correlation coefficient of the turbulent velocities at a given station? Another possibility of determining the size of the eddies is by measuring u'^2 as a function of the time interval over which the averaging is done.

Dr. H. Wexler (*U. S. Weather Bureau, Washington, D. C.*):

(1) If mechanical mixing changes the stratified water into homogeneous water, the mean temperature of the water should in either case be very nearly constant.

(2) If ozone sinks to the ground in polar regions because of the great stability of Arctic air, this sinking will occur immediately behind the moving cold front so that one might expect a maximum in surface-ozone content in a comparatively narrow belt (a few hundred kilometers wide) behind the cold front.

Dr. R. B. Montgomery (*New York University, New York, N. Y.*):

In regard to the autumnal temperature rise in the lowest levels of Lake Michigan at the time when tne temperature becomes nearly constant from top to bottom, it does not seem to be due necessarily to mechanical stirring produced by the wind. The depth involved, some 75 meters, is greater than experience indicates that wind stirring may commonly penetrate. An alternate possibility is to ascribe the mixing in the deep layers to the convection due to surface cooling. This interesting problem

deserves further study not only for its significance in fresh water, but also because a similar autumnal warming at some depth below the surface is observed to occur in the ocean.

Mr. Phillip Light (*U. S. Weather Bureau, Washington, D. C.*):

Is there a point of demarkation between logarithmic distribution of wind velocity at the foot of a slope and the type of distribution at a hill top?

REFERENCES

1. **Bagnold, R. A.**
1941. The physics of blown sand and desert dunes. London. Methuen & Co.

2. **Church, Phil E.**
1942. The annual temperature cycle of Lake Michigan, I. Cooling from late autumn to the terminal point, 1941–42. Misc. Repts. No. 4, Univ. of Chi., Inst. of Meteorol., Chicago, Aug. 1942, 50 pp. mimeo. 2 pl.

3. **Harrington, Mark**
1895. Surface currents of the Great Lakes. U. S. Weather Bureau Bull. **B**.

4. **Kármán, Th. von**
1934. Turbulence and skin friction. Jour. Aero. Sci. **1**: 1-20.

5. **Spilhaus, A. F.**
1937. A bathythermograph Jour. Marine Research **1**: 95–100.

6. 1940. A detailed study of the surface layers of the ocean. Jour Marine Research **3**: 51–75. See pp. 52–55.

7. **Wulf, O. R.**
1935. Light absorption in the atmosphere and its photochemistry. Jour. Optical Soc. Am. **25**: 231.

8. **Wulf, O. R., & Deming, L. S.**
1937. The distribution of atmospheric ozone in equilibrium with solar radiation and the rate of maintenance of the distribution. Jour. Terrestrial Magnetism and Elec. **42**: 195.

THE INFLUENCE OF STABILITY ON EVAPORATION

By Benjamin Holzman
U. S. Weather Bureau, Washington, D. C.

PART I

Richardson[5,6] and Prandtl[3,4] have shown that the production and dissipation of turbulent energy in a given air layer depends upon a critical value of a relationship between the lapse rate and the vertical wind-velocity distribution. The influence of stability was demonstrated to depend on the dimensionless number,

$$-\frac{g}{\rho}\frac{\partial \rho}{\partial z}\bigg/\left(\frac{\partial u}{\partial z}\right)^2 \qquad (1),$$

where g is the gravity term, ρ the density of the air, u the velocity and z the height. For the atmosphere this expression may be replaced by

$$\frac{g}{T}\left(\frac{\partial T}{\partial z}+\Gamma\right)\bigg/\left(\frac{\partial u}{\partial z}\right)^2 \qquad (2),$$

where T is the temperature and Γ is the adiabatic lapse rate equal to 1° C. per 100 meters. The quantity, $\frac{1}{\theta}\frac{\partial \theta}{\partial z}$, where θ is the potential temperature, may be substituted for the quantity

$$\frac{1}{T}\left(\frac{\partial T}{\partial z}+\Gamma\right)$$

and expression (2) becomes

$$\frac{1}{\theta}\frac{\partial \theta}{\partial z}\bigg/\left(\frac{\partial u}{\partial z}\right)^2 \qquad (3).$$

Thus, for adiabatic conditions, the stability equation becomes zero. When the atmosphere is thermally stratified, the expression is positive; when unstable, the equation becomes negative. When the expression reaches a certain positive value, turbulence tends to die out; when below this critical value, eddy energy is produced. There has been no agreement regarding the magnitude of this critical number. Richardson[6] estimated its value to be at 1, Prandtl[4] at ½ and Taylor[12] and Goldstein[1] at ¼. More recently, Schlichting[8] determined the critical number to be 1/24. Von Kármán[2] has suggested that the earlier determinations of the turbulence limit number are not strictly valid.

Stable stratification tends to damp out turbulent motion because part of the turbulent energy is used in performing the work of bringing potentially colder eddies upward or potentially warmer eddies downward. Using this principle Rossby and Montgomery[7] have extended the theory for turbulent transport in an adiabatic atmosphere to one of stable stratification.

Rossby points out that under conditions of stable stratification the mixing length is reduced. In addition, part of the turbulent kinetic energy is used up, due to the work that is done when the eddies are vertically displaced. Thus, if a vertical wind shear, $\partial u_s/\partial z$ or C_s, is assumed, in an adiabatic medium the turbulent kinetic energy would be proportional to $l^2C_s^2$. In a stable medium the mixing length would be reduced to l_s and the corresponding turbulent kinetic energy would be proportional to $l_s^2C_s^2$. The difference in the kinetic energies between the adiabatic and stable atmospheres must be essentially equal to the work that is done in the vertical displacement of eddies under stable conditions which can be shown to be equal to

$$\frac{g}{T} l_s^2 \left(\frac{\partial T}{\partial z} + \Gamma\right),$$

or in other words,

$$A l^2 C_s^2 = A l_s^2 C_s^2 + \frac{g}{T} l_s^2 \left(\frac{\partial T}{\partial z} + \Gamma\right) \qquad (4).$$

Factor A is a proportionality constant. It is not clear why the two A's should necessarily have the same value, but Rossby (1935 42) assumes they do and writes equation (4) as,

$$l^2 C_s^2 = l_s^2 C_s^2 + \frac{\beta g}{T} \left(\frac{\partial T}{\partial z} + \Gamma\right) l_s^2 \qquad (5),$$

where β is the proportionality factor. From the above equation it is evident that

$$l_s = \frac{l}{\sqrt{1 + \beta \frac{g}{T}\left(\frac{\partial T}{\partial z} + \Gamma\right) \Big/ \left(\frac{\partial u}{\partial z}\right)^2}} \qquad (6).$$

In transfer problems where stability must be considered, this expression (6) for the mixing length leads to a number of disagreeable integrals, depending on what assumptions are made regarding the lapse rate and wind shear. If, however, the expression for the mixing length under stable conditions were of the form,

$$l_s = l \sqrt{1 - \beta \frac{g}{T}\left(\frac{\partial T}{\partial z} + \Gamma\right) \Big/ \left(\frac{\partial u}{\partial z}\right)^2} \qquad (7),$$

a very convenient, integrable expression results for the vertical transport of meteorological properties under turbulent and stable conditions.

Rossby's stability correction for the mixing length requires that the stability number,

$$\frac{g}{T}\left(\frac{\partial T}{\partial z}+\Gamma\right)\bigg/\left(\frac{\partial u}{\partial z}\right)^2,$$

be equal to infinity before l_s becomes zero. As indicated previously it has been shown that when the stability number attains a certain critical value, turbulence is suppressed. This means that l_s must have a limiting value for the critical stability number and must rapidly approach zero when the critical stability number is exceeded.

Equation (7) rapidly approaches zero for l_s with increasing stability. Rossby's equation departs considerably from this curve with increasing stability and approaches the zero axis asymptotically. It should be pointed out, however, that Rossby's stability equation for the mixing length has a theoretical basis to recommend it. In any event both equations give approximately the same results for those stability numbers that are not too close to the critical stability number.

PART II

The derivation of a formula for evaporation when conditions of stability prevail can be developed in a manner similar to that under adiabatic conditions such as suggested by Rossby, Sverdrup, Montgomery and others.[7 10 11] Remembering that the mixing length[4] can be written as

$$l_s = k_0 z,$$

equation (7) can be written as

$$l_s = k_0 z\sqrt{1-\beta\frac{k^2}{\left(\frac{\partial u}{\partial z}\right)^2}} \qquad (8),$$

where

$$k^2 = \frac{g}{T}\left(\frac{\partial T}{\partial z}+\Gamma\right).$$

The so-called roughness parameter, z_0, has been purposely avoided because of its loose and inadequate definition.

Under thermal stratification, then,

$$\frac{\partial q}{\partial z} = -\frac{E}{\rho l_s \sqrt{\frac{\tau}{\rho}}} \tag{9}$$

or

$$\frac{\partial q}{\partial z} = -\frac{E}{\rho k_0 z} \cdot \frac{1}{\sqrt{1 - \beta \frac{k^2}{\left(\frac{\partial u}{\partial z}\right)^2}}} \cdot \frac{1}{\sqrt{\frac{\tau}{\rho}}} \tag{10}$$

where q is the specific humidity, τ is the shear stress, and E is the evaporation. Since

$$\frac{\partial u}{\partial z} = \frac{1}{l_s}\sqrt{\frac{\tau}{\rho}} \tag{11}$$

we can write

$$\frac{\partial u}{\partial z} = \frac{1}{k_0 z} \frac{\sqrt{\frac{\tau}{\rho}}}{\sqrt{1 - \beta \frac{k^2}{\left(\frac{\partial u}{\partial z}\right)^2}}} \tag{12}$$

Combining equations (10) and (12) and eliminating $\sqrt{\frac{\tau}{\rho}}$,

$$\frac{\partial q}{\partial z} = -\frac{E}{\rho k_0^2 z^2} \cdot \frac{1}{\left[1 - \frac{\beta k^2}{\left(\frac{\partial u}{\partial z}\right)^2}\right]} \cdot \frac{1}{\frac{\partial u}{\partial z}} \tag{13}$$

In the immediate vicinity of the ground a logarithmic distribution of properties may be assumed.[9, 10] Then if

$$\frac{\partial T}{\partial z} = \frac{a}{z}, \quad \frac{\partial u}{\partial z} = \frac{b}{z}.$$

Substituting these relations in equation (13),

$$\frac{\partial q}{\partial z} = -\frac{\frac{E}{\rho k_0^2 z b}}{1 - \beta \frac{\frac{g}{T}\left(\frac{a}{z} + \Gamma\right)}{\frac{b^2}{z^2}}} \tag{14}$$

The value of the adiabatic lapse rate, Γ, in the stability term is quite

small in comparison with the other terms in the numerator and can be neglected without any significant error. Simplifying equation (14),

$$\frac{\partial q}{\partial z} = -\frac{E}{\rho k_0^2 b} \cdot \frac{1}{z(1 - sz)} \tag{15}$$

where

$$s = \frac{\beta(g/T)a}{b^2}.$$

Equation (15) conveniently integrates into

$$q_1 - q_2 = \frac{E}{\rho k_0^2 b}\left[\ln\left(\frac{1 - sz_1}{z_1} \cdot \frac{z_2}{1 - sz_2}\right)\right] \tag{16}$$

But

$$b = \frac{u_2 - u_1}{\ln \frac{z_2}{z_1}}.$$

Then substituting for b and solving for E,

$$E = \frac{\rho k_0^2 (q_1 - q_2)(u_2 - u_1)}{\ln \frac{z_2}{z_1}}\left[\frac{1}{\ln\left(\frac{z_2}{z_1} \cdot \frac{1 - sz_1}{1 - sz_2}\right)}\right] \tag{17}$$

When adiabatic or near adiabatic conditions prevail in the atmosphere, the factor s is equal or nearly equal to zero and equation (17) becomes identical to the expression for evaporation under adiabatic conditions.[13] The parameter s contains a square of the wind shear in the denominator. This means that with any appreciable wind difference the numerical value of s is quite small and the corresponding influence of stability on evaporation will also be small.

The flow of stable air over rough surfaces, such as a vegetated or natural ground surface, is necessarily characterized by a rather steep vertical wind-velocity gradient which in turn would indicate that the influence of stability in suppressing evaporation can not be of great importance. It would appear that the only natural areas where the influence of stability might affect evaporation sufficiently to be of hydrologic significance would be fairly smooth-covered snow surfaces, under conditions of fairly quiescent air.

DISCUSSION OF THE PAPER

Prof. B. Haurwitz (*Massachusetts Institute of Technology, Cambridge, Mass.*):

Is it possible to apply the last formulas to mean values even though they have been derived for instantaneous data? It appears that it should be feasible to show that the evaporation equation holds also in the case of periodic changes such as the daily or yearly period, taking averages for the whole period and obtaining the average evaporation for the period.

Reply by Mr. Holzman:

The formula works very well for mean values, especially if the demand for accuracy is only of hydrologic significance. For example, hourly values of evaporation were computed from hourly observations of the moisture gradient and wind shear and these values were summed up over a 24-hour period and compared with a single daily (24-hour) computed evaporation from an average 24-hour moisture gradient and average 24-hour wind-shear observation. The agreement was quite good and certainly adequate for hydrologic purposes. But, for precision measurements, it is recognized that the use of mean values may not be suitable. This is true because a maximum moisture gradient may be associated with a minimum wind shear, or other combinations may occur so that an average value is not representative for the period.

REFERENCES

1. **Goldstein, S.**
   1931. On the stability of superposed streams of fluids of different densities. Roy. Soc. London, Proc. A, **132**: 524–548.
2. **Kármán, Th. von**
   1935. Some aspects of the turbulence problem. 4th Internat'l Cong. Appl. Mech. Cambridge, England, Proc. **1934**: 54–91.
3. **Prandtl, L.**
   1930. Einfluss stabilisierender Kräfte auf die Turbulenz. *In* Gilles, A., Hopf, L., and Kármán, Th. von, eds , Vortrage aus dem Gebiete der Aerodynamik und verwandte Gebiete (Aachen, **1929**): 1–7.
4. 1932. Meteorologische Anwendung der Strömungslehre. Beit. z. Phys. der freien Atmos. **19**: 188 202.
5. **Richardson, Lewis F.**
   1919a. Atmosphere stirring measured by precipitation. Roy. Soc. London, Proc. A, **96**: 9–18.
6. 1919b. The supply of energy from and to atmospheric eddies. Roy. Soc. London, Proc. A, **97**: 354–373.
7. **Rossby, C.-G., & Montgomery, R. B.**
   1935. The layer of frictional influence in wind and ocean currents. Mass. Inst. Technol. and Woods Hole Oceanographic Inst., Papers in Phys. Oceanography and Met. **3** (3): 101 p.
8. **Schlichting, H.**
   1935. Turbulenz bei Wärmeschichtung. 4th Internat'l Cong. Appl. Mech., Cambridge, England, Proc. **1934**: 245–246.
9. **Sutton, O. G.**
   1936–1937. The logarithmic law of wind structure near the ground. Roy. Met. Soc. (London), Quart. Jour. **62**: 124–127; **63**: 105–107.
10. **Sverdrup, H. U.**
    1936. The eddy conductivity of the air over a smooth snow field. Results of the Norwegian-Swedish Spitsbergen expedition in 1934. Geofys. Pub. **11** (7): 69 p.
11. 1939. On the influence of stability and instability on the wind profile and the eddy conductivity near the ground. 5th Internat'l Cong. Appl. Mech., Cambridge, Mass., Proc. **1938**: 369–372.
12. **Taylor, G. I.**
    1931. Effect of variation in density on the stability of superposed streams of fluid. Roy. Soc. London, Proc. A, **132**: 499–523.
13. **Thornthwaite, C. W., & Holzman, Benjamin**
    1939. The determination of evaporation from land and water surfaces. U. S. Monthly Weather Rev. **67** (1): 4–11.

SOURCES OF ATMOSPHERIC HEAT AND MOISTURE OVER THE NORTH PACIFIC AND NORTH ATLANTIC OCEANS*

By Woodrow C. Jacobs†

Scripps Institution of Oceanography, University of California, La Jolla, California

INTRODUCTION

The present paper is a summary of results of an investigation concerning the exchange of heat and water vapor between the sea surface and atmosphere over the North Pacific and North Atlantic oceans. The theoretical aspects of the investigation, together with the derivations of equations, have already been presented in a previous paper[2] and will not be elaborated here.

Briefly, the method consisted of developing an equation which would express the rate of evaporation as a function of the difference in vapor pressure between sea surface and atmosphere, and of the wind movement within the turbulent layer, using available climatic data over the oceans. The original Sverdrup equation[10] was used, assuming an atmosphere in unstable or neutral equilibrium, the constants of which were derived empirically by comparing the mean annual evaporation over the oceans as computed by such a formula with similar values obtained through the use of the energy equations. When using mean climatic data over the oceans, it was found that the evaporation could be determined by means of the following equation,

$$E = 0.143\,(e_w - e_a)\,U_a \qquad \text{mm/24-hours} \qquad (1),$$

where e_w is the vapor pressure at the sea surface (millibars), e_a the vapor pressure at 6 meters height and U_a the wind speed (meters per second) at 6 meters.

It was shown further that the amount of sensible heat exchanged between sea and atmosphere could be computed by means of the Bowen formula[1] in the form:

$$Q_c = 0.65\left(\frac{t_w - t_a}{e_w - e_a}\right) L_t\, E, \qquad (p = \text{const.} = 1000 \text{ mb}) \qquad (2),$$

and in which t_w and t_a are sea surface and air temperatures, respectively,

*Contributions from the Scripps Institution, New Series, No 201

†Present address, U. S. Weather Bureau, Washington, D C

L_t is the latent heat of vaporization of water at temperature t, and E is expressed in grams or centimeters.

Through the use of equations (1) and (2), computations of the mean seasonal values for evaporation, heat exchange and total energy exchange were made for each five-degree square in the North Pacific and North Atlantic. The charts constructed from the data showing the total energy exchange between sea and atmosphere are of particular interest to the field of oceanography since the loss or gain of energy from the sea surface through either conduction or evaporation is entirely in the form of heat. On the other hand, the charts showing the mean seasonal and annual values for E and Q_c are of interest to the meteorologist, who is concerned not only with the amount of heat added to or subtracted from the lower layers of the atmosphere, but also with the amount of latent energy in the form of water vapor which is rendered available for the various meteorological processes. The previously mentioned paper[2] emphasized the oceanographic aspects of the investigation; the present paper is more concerned with the meteorological significance of the results.

SEASONAL AND REGIONAL VARIATIONS IN THE RATE OF EVAPORATION OVER THE NORTH PACIFIC AND NORTH ATLANTIC

The distribution of values for the mean daily evaporation within each 5-degree square over the North Pacific and North Atlantic oceans, as computed by means of equation (1), during winter (December, January, February) and during summer (June, July, August) is illustrated in FIGURES 1 and 2. These charts show that evaporation over the oceans is considerably greater during winter than during the summer months, particularly at higher latitudes and over the western portions. They also show that, in general, the regions of greatest evaporation are located along the eastern coasts of Asia and North America over the Kuroshio and Gulf Stream. Secondary areas of high evaporation appear about the southern margins of the North Atlantic and North Pacific semi-permanent high-pressure cells, i.e., in the northern part of the trade-wind region. Thus, in general, it is shown that the regions of greatest evaporation are those within which the northerly transport of surface waters is greatest, and over areas that are subjected during winter to frequent invasions by cold, dry continental air masses from the interiors of Asia or North America. An examination of the mean fields of motion over the Far East and the eastern coasts of North America shows that the pre-

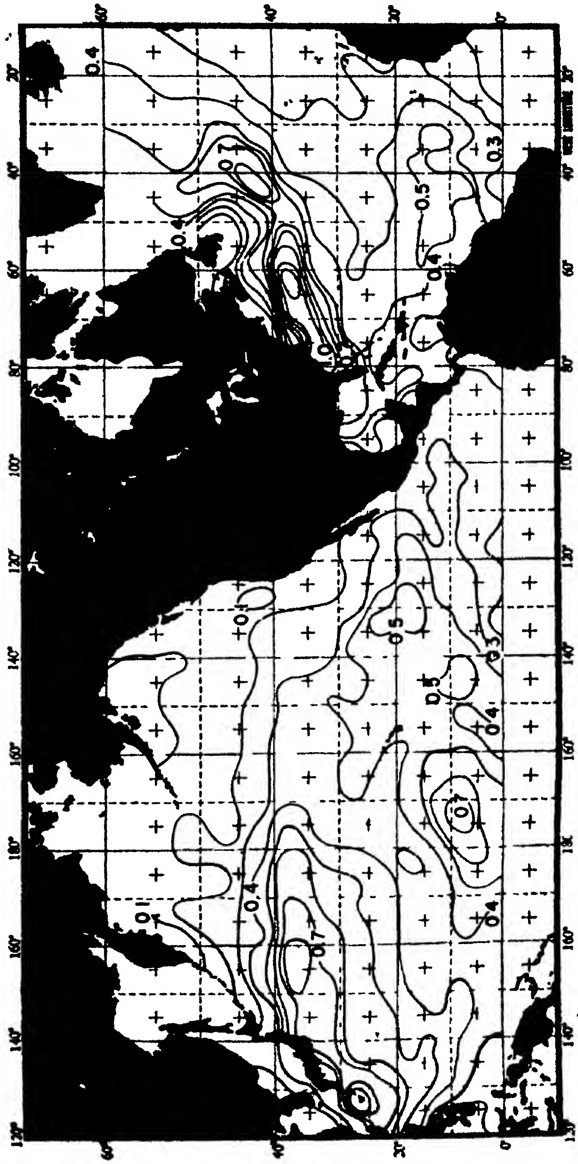


FIGURE 1.

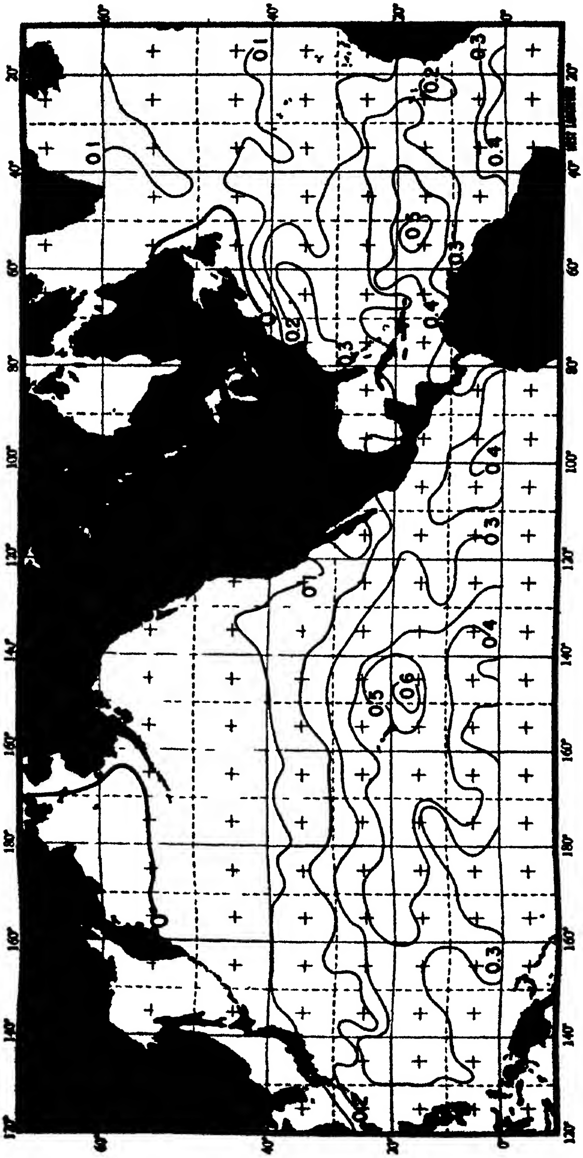


Figure 2

vailing air-mass types over the western sides of the oceans during winter are of continental origin, i.e., cPk or cAk air masses.*

The regions of least evaporation are those of southerly-flowing ocean currents along the extreme northwest coasts and along the eastern sides of the oceans. These are regions which, except in the northwest portions, are most frequently under the influence of air masses with a relatively long maritime history.

The great amount of evaporation within the trade-wind belts appears to be associated with the dry, descending air currents accompanying the semi-permanent fields of high pressure. A relatively low evaporation over the equator in the eastern North Pacific during spring appears to be associated with the northward-flowing ocean currents which in this region cross the equator resulting in an influx of cold water from higher latitudes in the Southern Hemisphere.

In summer the evaporation reaches its lowest value over most of the North Pacific and North Atlantic except within the eastern equatorial regions of both oceans, where the maximum values are reached during this season. A slight net condensation is computed for this season within the North Pacific north of latitude 55° and west of 160° W, and in the North Atlantic waters immediately surrounding Labrador.

In the North Atlantic Ocean, the center of the tropical or trade-wind area of maximum evaporation remains nearly stationary during all seasons except autumn, when it quite largely disappears, with its mean position located between latitudes 10° and 20° N, and approximately one-third the distance from the coast of South America to West Africa. In the North Pacific Ocean, on the other hand, this tropical area of high evaporation is considerably better developed and the position of its center varies with the seasons. In winter the maximum evaporation within this area is 0.70 cm/day and the center is located between 5° and 10° N latitude and 170° and 175° W longitude. In spring the maximum increases slightly to 0.72 cm/day and the center is displaced northeastward to the area between 10° and 15° N latitude and 150° and 155° W longitude. In summer the maximum evaporation in this region decreases to 0.62 cm/day and the center is displaced farther northeastward to the area between 15° and 20° N latitude and 145° and 150° W longitude. In the autumn, however, the maximum evaporation within this area falls to its lowest value, 0.56 cm/day, and the now less well-defined center is displaced southward to the general region on the equator between 0° and 5° N latitude, and 135° and 150° W longitude.

**Note:* cPk = Polar continental air colder than the surface over which it is passing. cAk = Arctic air colder than the surface. For description of the air-mass types see Petterssen, Weather Analysis and Forecasting, pp. 166–204 (1941).

This clockwise migration of the tropical center of high evaporation in the North Pacific appears to correspond to the similar movement of the center of the North Pacific high-pressure field.

The average seasonal values of evaporation for the various latitude ranges are given in FIGURE 3. These curves illustrate very clearly the

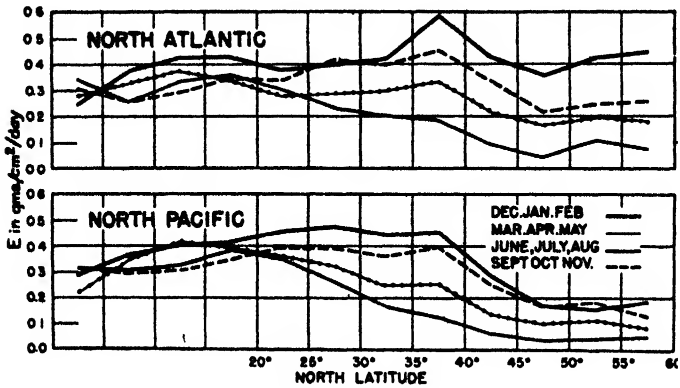


FIGURE 3.

winter maximum of evaporation in nearly all latitudes and the much lower summer evaporation within the higher latitudes. They also show that the seasonal variation in evaporation at low latitudes is much smaller than in the higher latitudes; that evaporation is higher during autumn than during spring in the middle and high latitudes but that, except at the equator, the reverse is true at the lower latitudes. Within the latitude range 0° to 5° N in the North Atlantic, the evaporation values are lowest during winter but in the North Pacific they are lowest during spring.

THE TOTAL EVAPORATION OVER THE NORTH PACIFIC AND NORTH ATLANTIC

As shown in FIGURE 4, the latitudinal distribution of values representing the total volume of water evaporated from the oceans (ΣE) is quite different from the distribution of values representing the evaporation per unit area of sea surface (E). This is due to the unequal distribution of areas of sea surface over the Northern Hemisphere (see TABLE 1). These curves show that the greatest quantity of water is evaporated in

the lower latitudes during all seasons and that the quantity is greatest in winter, least in summer and somewhat greater during autumn than during spring.

As shown by the data in TABLE 2, the seasonal variations in ΣE are greater in the North Atlantic than in the North Pacific, the percentage difference between the winter and summer evaporation being 13.8 per cent of the total annual evaporation in the North Atlantic and only 9.6 per cent of the total annual evaporation in the North Pacific. Although

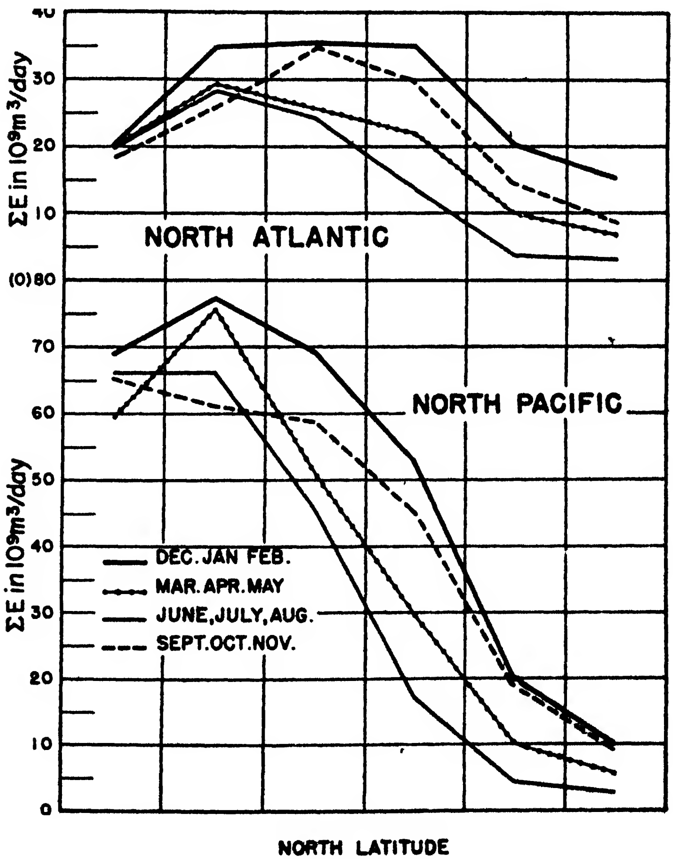


FIGURE 4.

TABLE 1

DISTRIBUTION OF AREAS IN THE NORTH PACIFIC AND NORTH ATLANTIC BY LATITUDE ZONES AND AS A WHOLE*

North Latitude range (°)	North Pacific: Area 1000 km²	North Pacific: Per cent of total (North Pacific plus North Atlantic)	North Pacific: Per cent of total (North plus South Pacific)	North Pacific: Per cent of North Pacific	North Atlantic: Area 1000 km²	North Atlantic: Per cent of total (North Pacific plus North Atlantic)	North Atlantic: Per cent of total (North plus South Atlantic)	North Atlantic: Per cent of North Atlantic
0–10	21,246	76.4	12.0	26.0	6,545	23.6	7.2	14.3
10–20	18,656	69.5	10.5	22.9	8,181	30.5	9.0	17.9
20–30	14,922	61.9	8.5	18.3	9,180	38.1	10.1	20.0
30–40	11,707	62.6	6.6	14.3	7,000	37.4	7.7	15.3
40–50	8,685	62.6	4.8	10.6	5,181	37.4	5.7	11.3
50–60	5,773	62.0	3.3	7.1	3,545	38.0	3.9	7.7
(60–70)†	(643)	(17.6)	(0.4)	(0.8)	(3,000)	(82.4)	(3.3)	(6.5)
(>70)	—	(0.0)	—	—	(3,182)	(100.0)	(3.5)	(7.0)
	Total 81,632	% of total 64.1	% of total 46.1	Total 100.0	Total 45,814	% of total 35.9	% of total 50.4	Total 100.0

*Computed from data by Schott.[7,8]

†Areas in parentheses are not included in the energy computations.

the area of the North Pacific between the equator and latitude 60° N ($80,989 \times 10^3$ km²) is 67.1 per cent of the total area ($120,621 \times 10^3$ km²) of both oceans between the same parallels, it accounts for only 66.5 per cent of the total volume of water evaporated from both areas, due to the slightly greater rate of evaporation in the North Atlantic. The total volume of water evaporated from the North Pacific during the course of the year averages $90,232.4 \times 10^9$ m³, which is equivalent to an evaporation rate of 111.4 cm/year. The total volume of water evaporated from the North Atlantic during the year amounts to $45,490.0 \times 10^9$ m³, which is equivalent to a rate of 114.8 cm/year. The quantity for both oceans is $135,722.4 \times 10^9$ m³/year, equivalent to a mean depth of 112.5 cm/year. The higher rate of evaporation in the North Atlantic appears to be due largely to the smaller amount of sensible heat given off to the atmosphere rather than to any significant excess of energy available at the surface. But an analysis of cloud charts for the Northern Hemisphere[9] indicates that the mean annual cloudiness between the equator

TABLE 2

MEANS OF THE TOTAL DAILY EVAPORATION (ΣE) OVER THE NORTH PACIFIC AND NORTH ATLANTIC

Area	Season	Mean ΣE	ΣE for both oceans	Mean annual ΣE
		10^6 m³/day	Per cent	Per cent
North Pacific	Winter	297,019	64.8	30.0
	Spring	231,834	67.3	23.5
	Summer	202,054	68.5	20.4
	Autumn	257,942	66.2	26.1
	Year	247,212	66.5	100.0
North Atlantic	Winter	161,314	35.2	32.4
	Spring	112,637	32.7	22 6
	Summer	92,940	31.5	18.6
	Autumn	131,627	33.8	26 4
	Year	124,630	33.5	100.0
Both Oceans	Winter	458,333	100.0	30.8
	Spring	344,471	100 0	23.2
	Summer	294,994	100.0	19.8
	Autumn	389,569	100.0	26.2
	Year	371,842	100.0	100.0

Total Annual Evaporation:—

North Pacific
= $90,232.4 \times 10^9$ m³/year: equivalent to mean depth of 111.4 cm/year.

North Atlantic
= $45,490.0 \times 10^9$ m³/year: equivalent to mean depth of 114.8 cm/year.

Both oceans
= $135,722.4 \times 10^9$ m³/year: equivalent to mean depth of 112.5 cm/year.

and latitude 60° N may be somewhat greater in the North Pacific than in the North Atlantic. Rough estimates by scalage indicate the average annual cloudiness over the North Pacific is about 60 per cent; over the North Atlantic about 55 per cent. If these latter figures are valid, they may partly account for the greater evaporation rate in the North Atlantic.

By means of reductions from pan observations at sea, Wüst (1920; 1936) has derived values for the mean annual evaporation at various latitudes in the several oceans. His figures for the North Pacific and North Atlantic, however, extend only from the equator to latitude 40° N. By taking the mean of Wüst's values for three adjacent 5-degree parallels as being representative of the mean evaporation within each 10-degree latitude zone, his figures show an evaporation of 127.5 cm/year in the North Atlantic and 118.3 cm/year in the North Pacific. The present calculations for the same latitude range (0° to 40° N) give 123.0 cm/year for the North Atlantic and 124.5 cm/year for the North Pacific. The mean value for this same area in both oceans is 124.0 cm/year, thus about 2 per cent greater than Wüst's average (121.2 cm/year) for the two oceans.

Wust gives 93 cm/year as the average evaporation for all oceans This low value is due to including evaporation in the Southern Hemisphere where the values appear to be lower at most latitudes than in the Northern Hemisphere. This latter condition is particularly true over the extensive areas of sea surface at high latitudes in the Southern Hemisphere. For example, his value for the latitude range (45° to 50° S) within the Atlantic, Pacific and Indian Oceans is only 43 cm/year.

The seasonal ratios between the total daily evaporation in the North Pacific and North Atlantic are given in TABLE 3. These data show the ratio $\Sigma E_{\text{Pacific}}/\Sigma E_{\text{Atlantic}}$ to be greatest during summer, least during winter and greater during spring than autumn. The reasons for these variations appear to be greater contrasts between the winter and summer air-mass types over the North Atlantic than is the case in the North Pacific. More explicitly, the winter air masses over the North Atlantic are drier and colder relative to the summer air masses than are those of the North Pacific, which is not surprising when one considers the smaller area of the North Atlantic and the corresponding closer proximity of the continental land masses with their wide seasonal variations in temperature and humidity.

It was brought out in the previous paper that the locations of the principal frontal zones over the oceans as given by Petterssen[6] correspond quite closely with the zones of maximum energy exchange. The trade-wind area of maximum evaporation in the North Pacific lies im-

TABLE 3

SEASONAL RATIOS BETWEEN TOTAL DAILY EVAPORATION IN THE NORTH PACIFIC AND ATLANTIC OCEANS

Season	Ratio $\frac{\Sigma E \text{ (N. Pacific)}}{\Sigma E \text{ (N. Atlantic)}}$
Winter	1.84
Spring	2.06
Summer	2.17
Autumn	1.96
Year	1.98

mediately east and south of his mean winter position for the mid-Pacific Polar Front in the Northern Hemisphere, with the principal axis of the frontal zone parallel to the major axis of the zone of maximum evaporation. On the other hand, the maximum zones of winter evaporation in the western parts of the two oceans lie largely to the west and north of Petterssen's positions for the polar fronts in winter over these areas. The lines or zones of frontogenesis (cyclogenesis), however, will coincide more nearly with the zones of maximum temperature gradient immediately off the coasts of East Asia and eastern North America. Thus, in the western portions of the oceans, the principal frontogenetic areas will be located along the eastern coasts of the two continents northwestward from the mean positions of the polar fronts themselves. This is due to the fact that the zones of frontogenesis will tend to move with the air currents toward the axis of outflow and will assume a mean position near this axis or trough line in the mean flow. It is found that in both oceans the zones of frontogenesis (cyclogenesis) will lie in every case to the northwestward of a zone of maximum evaporation and that the major axes of the zones of maximum evaporation will parallel the principal axes of the frontal zones.

It is to be remarked that the mean seasonal data shown in FIGURES 1 and 2 are by no means representative of instantaneous conditions over the oceans. It is expected that non-periodic variations in the rates of evaporation over the oceans will occur and that the areas of maximum and minimum evaporation will also vary in both intensity and position. Thus over the Kuroshio and Gulf Stream, at the times of outbreaks of cold cAk and cPk air from the continents in winter, the evaporation values may occasionally be extremely high. Conversely, during periods in winter when the polar fronts are located near the coasts and the air

masses over the neighboring ocean areas are of southerly origin (mE or mT masses),* the evaporation in these areas must reach low values. Also, along the west coasts of the continents, the evaporation in winter may reach high values at the times of outbreak of dry air from the interior, as during the periods when the so-called "Santa Ana" winds blow from the east or northeast along the coasts of Southern California. Just as there are important non-periodic variations in the positions of the principal centers of action in the atmosphere, the polar fronts, etc., with the resulting time or regional variations in humidity, wind, temperature, cloudiness, etc., so must there be corresponding changes in the evaporation over the oceans. It is expected that the regional variations in evaporation (or heat exchange) will be most pronounced during periods with a low zonal index and that the evaporation over the oceans will be most uniform during periods with a high zonal index.[5]

If secular variations in the amount of solar energy received at the surface of the earth occur, it is to be expected that similar secular variations in evaporation and heat exchange will also take place. But it is not suggested that such variations must be proportional, since the effects of increased humidity, extent of cloudiness, etc., may be important.

VARIATIONS IN THE AMOUNT OF SENSIBLE HEAT EXCHANGED BETWEEN SEA AND ATMOSPHERE

The distribution of values representing the amount of sensible heat exchanged between sea and atmosphere over the North Pacific and North Atlantic during winter and summer are shown in FIGURES 5 and 6. The isolines for Q_c show, roughly, the same configuration as shown for E, these values being at their maximum in winter and along the western sides of the oceans. One very important difference is shown. No tropical areas of maximum heat exchange exist within the trade-wind region to correspond to the areas of maximum evaporation that appear to be associated with the North Pacific and Azores high-pressure fields.

During winter there are three areas of maximum heat exchange in the North Pacific, the principal area occurring within the Kuroshio between 35° and 40° N, and 150° and 155° E where Q_c averages 246 g.cal.cm^2/day, with secondary areas northeast of Formosa (Taiwan) and in the northern portion of the Sea of Japan. The area northeast of Formosa coincides with the area of maximum winter evaporation in the North Pacific (see FIGURE 1). The maximum value for Q_c in the North Atlantic (274 g.cal.cm^2/day) occurs within the Gulf Stream between latitudes 35° and

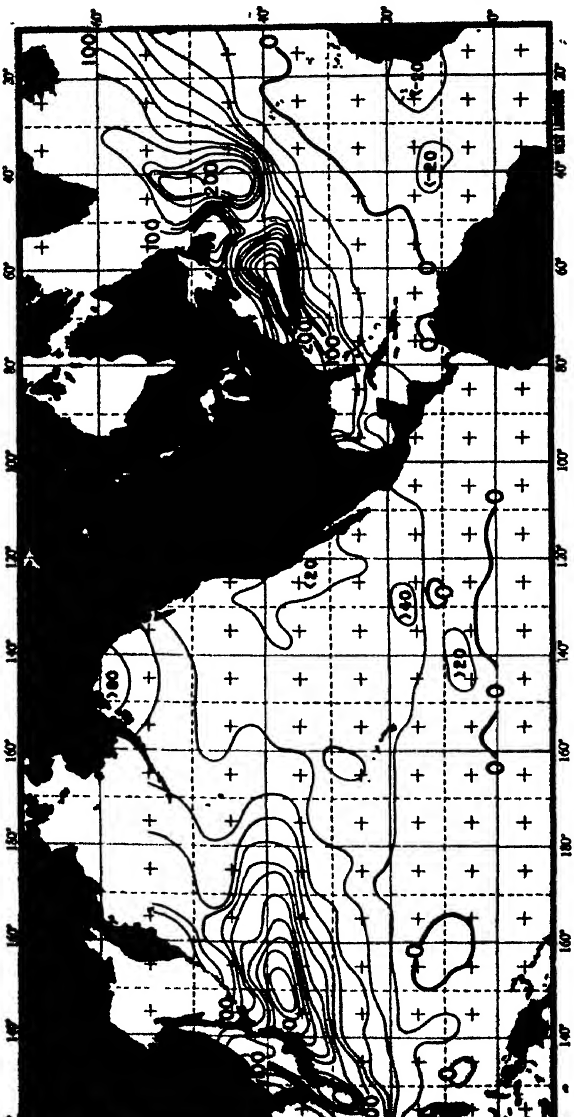
*mE = maritime equatorial air; mT = maritime tropical air.

40° N, about 700 miles east of the Virginia capes, with a secondary belt of high values for Q_e extending north and south along the 40° W meridian between latitudes 40° and 55° N. The North Pacific values for Q_e during winter are positive for all regions except for several small areas near the equator. In the North Atlantic, however, the charts show that Q_e is negative during winter for approximately one-third the area of that ocean.

In summer the values for Q_e become insignificant nearly everywhere in the North Pacific except in the southeastern equatorial regions where values as high as 40 g.cal.cm²/day are determined, and are negative everywhere north of latitude 40°, off the California coast and, surprisingly, also off the Asiatic coasts south to the Philippine Islands. In the North Atlantic during summer, Q_e is negative everywhere except within a narrow zone along the eastern coasts of North America extending northeastward from Florida to latitude 40° N, and within the Gulf of Mexico and Caribbean Sea. The average seasonal values of Q_e for the different latitude ranges are given in FIGURE 7. These curves illustrate very clearly the winter maximum in the heat exchange at nearly all latitudes and the pronounced maxima for Q_e above latitude 35° N in both oceans. During autumn, winter and spring, the values for Q_e increase generally from the equator to maxima in both oceans between 35° and 40° N, decrease to minima between latitudes 45° and 50° N, and then increase again poleward. During summer, however, the values for Q_e in both oceans generally decrease from the equator toward higher latitudes. One interesting point to note in these curves is the fact that the values for Q_e are higher in the North Pacific than in the North Atlantic at all latitudes and during all seasons except during winter for the areas north of latitude 35° N. This question will be discussed in greater detail under the section on the regional and seasonal variations in the values for the ratio between evaporation and heat exchange.

From the data so far presented it thus appears that the sea is heating the atmosphere by significant amounts only in the middle and high latitudes, along the eastern sides of the continents and principally during the winter season. In other portions of the oceans and almost everywhere during summer, the sea is actually receiving some energy from the atmosphere.

Data showing the latitudinal distribution of the mean annual values for Q_e indicate that in both oceans the maxima occur between latitudes 35° and 40° N, with low values at the equator and a secondary minimum in each ocean between latitudes 45° and 50° N. It has previously been pointed out that the mean annual rate of evaporation in the North



Fi

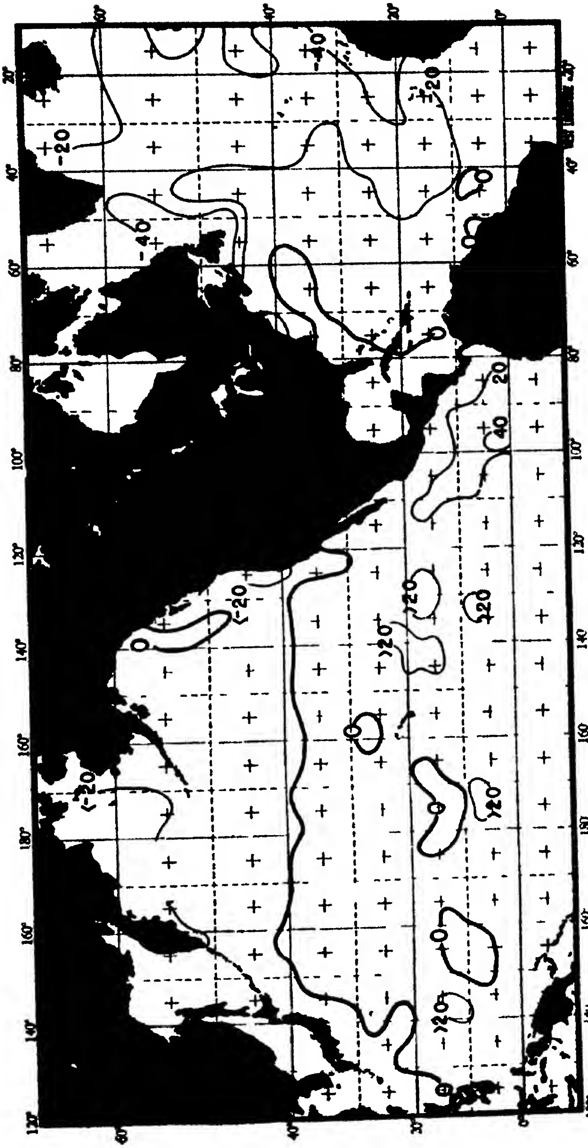


FIGURE 6

Pacific is highest between latitudes 20° and 25° N and in the North Atlantic between latitudes 35° and 40° N, with a secondary maximum

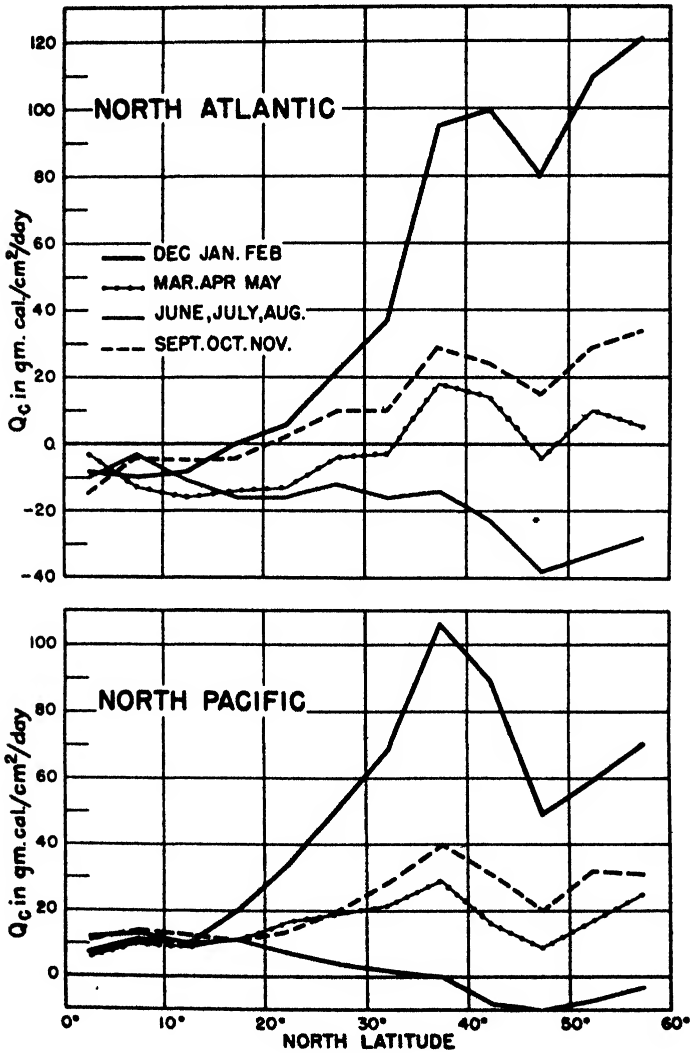


FIGURE 7.

between latitudes 15° and 20° N. Comparing the two sets of data, it is found that the zone of maximum evaporation in the North Pacific occurs farther south than the zone of maximum heat exchange, the displacement being 15 degrees of latitude. In the North Atlantic, however, the zones of maximum evaporation and maximum heat exchange coincide except that there is no secondary zone of maximum heat exchange within the latitude range 15° and 20° N to correspond to the secondary zone of maximum evaporation in this area. These conditions indicate that over the oceans as a whole the atmosphere is being heated most rapidly in the higher latitudes but that it is receiving moisture principally in the middle and lower latitudes.

VARIATIONS IN THE RATIO (R) BETWEEN THE AMOUNT OF HEAT GIVEN OFF TO THE ATMOSPHERE AS SENSIBLE HEAT (Q_s) AND THE AMOUNT OF HEAT USED FOR EVAPORATION (Q_e)

In previous investigations of evaporation from the oceans, the ratio between the amount of heat given off to the atmosphere by convection and the amount of heat used for evaporation was assumed a constant for lack of better information. For example, in their computations of the mean annual evaporation at various latitudes over the oceans, Mosby[4] has assumed R constant at 0.10, McEwen[3] at 0.20. However, the present computations show that this ratio is a highly variable quantity, both seasonally and with respect to the regional distribution. The mean seasonal values for R arranged by latitudes are shown in FIGURE 8. These curves illustrate that at all latitudes, the values for R are higher in the North Pacific than in the North Atlantic; that they are highest at all latitudes in both oceans during winter, lowest during summer, and about the same in spring as in autumn. During all seasons except summer, the values for R in both oceans increase from low values at the equator to maxima at high latitudes. During summer, the values are highest near the equator and decrease to minima in both oceans between latitudes 45° and 50° N, and then increase toward the poles. Thus (except during summer), while evaporation within the equatorial and tropical regions is high, the exchange of heat between sea and atmosphere is relatively small and, conversely, although evaporation at high latitudes is small, the exchange of heat between sea and atmosphere is relatively great. Similarly, it must be concluded that a greater amount of the available energy at the sea surface in the North Pacific is used in heating the atmosphere directly than is the case in the North Atlantic where a greater proportion is used in the evaporation process.

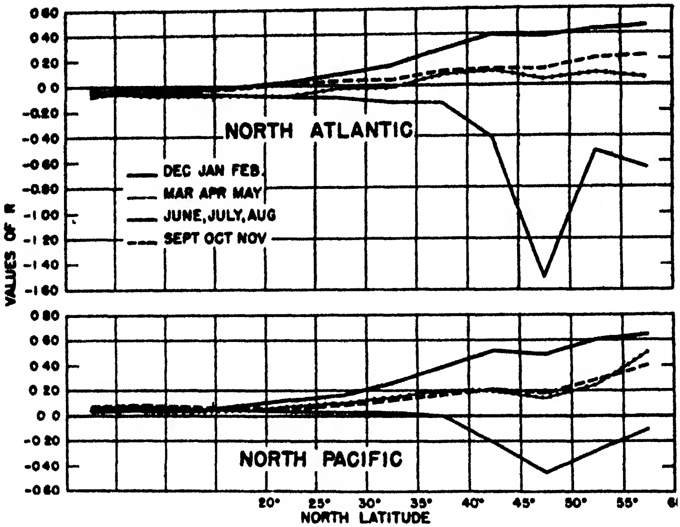


FIGURE 8.

The principal reason for the higher values for R in the North Pacific appear quite obvious. Since the ratio, Q_e/Q_e, is given by

$$R = 0.65 \frac{(t_w - t_a)}{(e_w - e_a)} \frac{p}{1000} \tag{3},$$

it follows that either vapor pressures over the North Pacific must be lower than over the North Atlantic or the quantities $(t_w - t_a)$ must be greater. The data show no significant differences between vapor pressures over the two oceans but an analysis of charts showing the differences between sea and air temperatures over the North Pacific and North Atlantic (U. S. Weather Bureau, 1938) shows very clearly that the positive temperature differences (sea minus air) are greater over the North Pacific as a whole than over the North Atlantic. The curves showing the latitudinal distribution of the mean annual values for R indicate, however, that the differences (R_{Pacific} minus R_{Atlantic}) are very nearly constant for all latitudes at about 0.10 or 0.11 (see TABLE 4). It is difficult to account for this constancy of differences.

It is to be noted that all computations of evaporation, heat exchange and total energy exchange have been made from data accumulated over the two oceans at or near Greenwich mean noon. Thus, most of the observations in the North Atlantic have been made during daylight hours

TABLE 4
LATITUDE VARIATION OF DIFFERENCE BETWEEN R OF THE PACIFIC AND OF THE ATLANTIC

North Latitude Range (°)	*R (Pacific) Minus R (Atlantic)	
0–5	0.10	
5–10	0.11	(0.106)
10–15	0.10	
15–20	0.10	
20–25	0.11	(0.106)
25–30	0.08	(0.11)
30–35	0.14	
35–40	0.10	
40–45	0.11	
45–50	0.12	
50–55	0.10	
(55–60)†	(0.21)	
Mean	0.11	

*Mean of the ratio between the average annual values for Q_e and Q_s.
†Not considered in the mean.

between 0600 and 1200, local mean time, while those in the North Pacific have been made largely during night hours between 2000 and 0600, local mean time. Although the diurnal variations in evaporation or heat exchange over the oceans can be assumed to be small (Sverdrup, 1940), the values computed from the uncorrected observational data obtained on shipboard might conceivably lead to significant differences and thus might account for the constancy of differences between the values for R in the North Pacific and North Atlantic. The hourly observational data necessary for computations of the hourly evaporation or heat exchange over the oceans are not generally available. There is a long series of observations for the area around the East Indies (between latitudes 12° S and 10° N; longitudes 112° and 134° E) obtained by the *Snellius Expedition*.[13] The results of the computations of R on the basis of the bihourly values of sea-surface temperature, air temperature, vapor pressure and wind speed for areas farther than 100 km. from the coasts, are shown in FIGURE 9. The curve showing the diurnal variation of R is rather irregular but with a decided minimum at midnight and a well-defined maximum at about 1000. Actually, the variations are of small magnitude since the vertical scale on FIGURE 9 is greatly exaggerated. Between the hours of 0600 and 1200, the mean value for R proves to be 0.070; between the hours of 2000 and 0600, 0.052. Thus it appears that

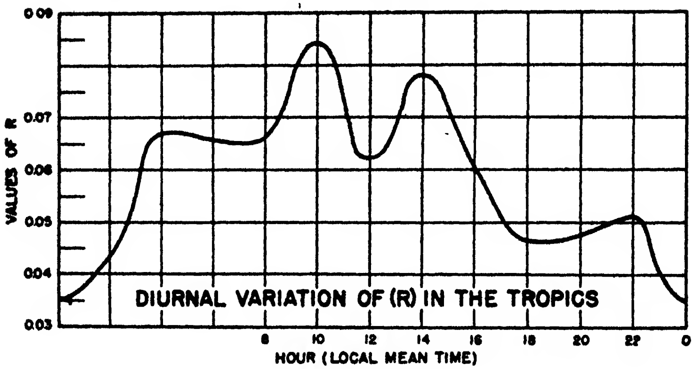


FIGURE 9.

the constancy of differences between the computed values for R in the North Pacific and North Atlantic at the various latitudes can hardly be explained on the basis of the diurnal variabilities in the quantities Q_c and Q_e, and the reasons for the constancy of differences are not apparent at the present time.

Before closing the discussion of the seasonal and latitudinal variations in R, it should be brought out that the computed values at very high latitudes may be questionable. This is true for two reasons: (1) the areas involved in the computations of R for high latitudes are rather small; and (2) equation (3) is very sensitive to small, and otherwise insignificant, observational errors (or the dropping of decimals in the means) at very low values for the quantities $(e_w - e_a)$ or $(t_w - t_a)$. Thus, at high latitudes (above 55° N) where Q_c and Q_e are very small, these small errors may be significant when computing R.

CONCLUSION

The present paper has attempted merely to present a few of the more important conclusions to be drawn from the computations of mean seasonal evaporation and heat exchange over the North Pacific and North Atlantic. The investigation is being continued at the Scripps Institution with some emphasis being placed upon the short-period and non-periodic variations in these quantities over the oceans. It is expected that these results, together with the complete series of charts showing the seasonal and annual values for E, Q_c and Q_e (the total energy exchanged between sea and atmosphere $= Q_e + Q_c$) and more detailed analyses of the data, will be presented in a later publication.[2]

ACKNOWLEDGMENT

The writer wishes gratefully to acknowledge the assistance and advice given by Dr. H. U. Sverdrup of the Scripps Institution of Oceanography, during all stages of the investigation.

DISCUSSION OF THE PAPER

Prof. C. F. Brooks (*Blue Hill Observatory, Milton, Mass.*):

I wonder if the systematic error in the determination of sea-surface temperature by bucket—the method used in the observations summarized and mapped by the Weather Bureau—may not be involved to a significant degree in the difference in computed evaporation between the Atlantic and Pacific in the diurnal range of R and in the general total. In 5 observational periods totalling 9 weeks at sea, both in the Atlantic and Pacific, in all seasons of the year and over a range of latitude from 7° to 60°, I have checked the accuracy of sea-temperature, air-temperature and humidity observations made by ships' personnel on American, Canadian, German and Norwegian steamships—with some surprising results. Only the sea-temperature checks have been published (in part).* The systematic error of a determination of sea-surface temperature is of the order of $-\frac{1}{2}°$ C., the observed temperature being below the true temperature. In the Gulf Stream off Cape Hatteras in winter the average error is about $-3°$ C.

The error at night is greater than by day, by perhaps $\frac{1}{2}°$ C. on the general average, but increasing in proportion to the evaporation. At night: (1) the wet bucket is colder, not being sunned; (2) the making of a catch takes more time; (3) the reading of the thermometer takes longer—either bucket or thermometer must be taken to a light. Thus, the computed evaporation during the night in the Pacific should be lower than during the day in the Atlantic Furthermore, the error of the bucket reading will be proportional to the evaporation. Therefore, the ratio R will differ by essentially the same amount in all latitudes, as found by Mr. Jacobs.

The diurnal course of R is apparently also influenced by this diurnal course of error.

The total annual computed evaporation is probably below the true evaporation. Perhaps this explains why Mr. Jacobs' values were somewhat lower than previous computations.

Observations of air and wet-bulb temperatures are subject also to appreciable error. Air temperatures reported at sea in the daytime are almost universally too high—owing to the ship's influence. Wet-bulb temperatures are also too high, and for added reasons. The wet-bulb covering is usually saline. (The wet-bulb is sometimes too high by being immersed in a bottle of water!) The result is that the vapor pressures computed from observations at sea are too high.

Reply by Mr. W. C. Jacobs:

It was realized in the beginning that the observational data over the oceans would present serious errors. However, the final evaporation equation obtained was prepared to fit just these sort of climatic data. Such a procedure is valid if the errors of observation can be handled in statistical fashion - the computations show that such is the case, therefore we cease to have any further interest in their magnitude (see ref. 2 for full discussion).

Prof. Brooks suggests that the evaporation computed by the above method may prove to be too low. While this might be the case when applying the theoretical Sverdrup equation to raw ocean data, the assigning of correction factors based on comparisons between annual evaporation amounts computed on the basis of the several energy equations and those computed on the basis of the Sverdrup equation has raised the amounts by approximately 30 per cent. Contrary to the statement of Prof. Brooks, the computed values are not lower than those of previous investigators. For example, Wüst's annual values are 2 per cent lower for both

*C. F. Brooks, Mon. Weather Rev., 54: 241, 1926; Journal Wash. Acad. Sci., 18: 529, 1928.

oceans (3 per cent lower in the case of the North Pacific; no difference in the case of the North Atlantic). Similarly, the computed evaporation is slightly higher than that given by McEwen and by Schmidt but slightly lower than that of Mosby, the latter of whom admits that his values constitute an upper limit and are probably too high to accurately represent average conditions. The lower evaporation rate in the North Pacific is also in agreement with previous investigations.

FIGURE 9 was prepared to show the small magnitude of the differences in the computed values for E and Q_e resulting from time differences in observations and was not intended to show the true diurnal variability of R. Correcting the data for diurnal effects (or errors), then, would actually result in lower values for R (higher evaporation) in the North Atlantic and higher values for R (lower evaporation) in the North Pacific, thus increasing the constant range of differences shown in TABLE 4 instead of decreasing them as is suggested by Prof. Brooks. Actually, however, the total difference (0.018) is too small to be considered of great significance.

REFERENCES

1. **Bowen, I. S.**
   1926. The ratio of heat losses by conduction and by evaporation from any water surface. Phys. Rev. **27**: 779-787.
2. **Jacobs, W. C.**
   1942. On the energy exchange between sea and atmosphere, Sears Found Jour. Mar. Res. **5** (1): 37-66.
3. **McEwen, George M.**
   1938. Some energy relations between the sea surface and the atmosphere. Sears Found. Jour. Mar. Res. **1** (3): 217-238.
4. **Mosby, H.**
   1936. Verdunstung und Strahlung auf dem Meere. Ann. der Hydrog. u. Marit. Meteor. **64**: 281-286.
5. **Namias, J., & Wexler, H.**
   1942. Suggested in a personal communication to the writer.
6. **Petterssen, Sverre**
   1940. Weather analysis and forecasting. 1st ed. New York. 503 p.
7. **Schott, Gerhard**
   1912. Geographie des Atlantischen Ozeans. Hamburg. 330 p.
8. 1935. Geographie des Indischen und Stillen Ozeans. Hamburg. 413 p.
9. **Shaw, Sir Napier**
   1936. Manual of meteorology. Vol. II. 2d edition. London. p. 146.
10. **Sverdrup, H. U.**
    1937. On the evaporation from the oceans. Sears. Found. Jour. Mar. Res. **1** (1): 3-14.
11. 1940. On the annual and diurnal variation of the evaporation from the oceans. Sears. Found. Jour. Mar. Res. **3** (2); 93-104.
12. **U. S. Weather Bureau**
    1938. Atlas of climatic charts of the oceans.
13. **Visser, S. W.**
    1936. The Snellius-Expedition. Vol. III, Meteorological results. Leiden. 111 p.
14. **Wüst, G.**
    1920. Die Verdunstung auf dem Meere. Veröff. Inst. f. Meereskunde, Berlin. N. F., A, Geogr. Naturw. Reihe, Heft 6, Oct. 1920.
15. 1936. Oberflächensalzgehalt, Verdunstung, und Niederschlag auf dem Weltmeere. *In:* Festschrift Norbert Krebs. 347-359.

TURBULENCE AND THE TRANSPORT OF SAND AND SILT BY WIND

By A. A. Kalinske

Iowa Institute of Hydraulic Research, State University of Iowa, Iowa City, Iowa

INTRODUCTION

Experimenters have recognized what appear to be three distinct methods of sand and silt transport by moving fluids: surface creep (called bed-load movement in the literature of hydraulics), saltation, and suspension. The second method has been studied and distinctly observed only in air, though, of course, there is no reason to suppose it does not occur in water.

The above classification seems to be fundamentally sound because each type of sand movement is controlled by a different set of forces acting in the fluid medium. When the drag forces on a sand particle overcome the gravity forces, the particle is displaced and rolls along with the fluid. The fluid drag on a grain can be expressed as

$$f = C\rho d^2 U^2 \qquad (1),$$

where d = particle size,
U = fluid velocity at particle,
C = coefficient depending on Reynolds number of grain, and
ρ = fluid density.

In rolling along, some of the grains get lifted slightly off the bed and thus get into the faster moving fluid, thereby acquiring additional energy. If this energy is sufficiently high the grain on dropping down will impart some of this energy to another grain, knocking it off the bed; or perhaps it bounces back itself. Thus grains are carried along in a region near the bed at a speed corresponding to the moving fluid. Such transport of sand is called *saltation* and has been vividly described by Bagnold[1] as it occurs in the desert during high winds.

At still greater velocities sand and silt particles may be kept in the moving fluid stream for prolonged periods by upward velocity components due to the turbulence, these upward components being at times greater than the velocity of fall of the particles. This mode of transportation is called *suspension*.

It appears that saltation is relatively unimportant in the transport of sediment in rivers. This is readily explained, since for an equal drag

force on any given size of particle the velocity in air will have to be of the order of $\sqrt{\rho_w/\rho_a}$ or about 30 times as great as in water. Thus, once particles are moving in air they will acquire from the air stream a momentum 30 times greater than particles under similar conditions would acquire in water. In fact, as Bagnold[1] points out, once saltation begins in air the drag between the moving air stream and the boundary is transmitted almost entirely through the action of the saltating particles and not by direct action of the fluid on the boundary. This action profoundly alters the velocity distribution in the region near the bed.

The criterion as to whether saltation is significant or not under any given conditions will be developed further in this paper. Also, an analysis will be indicated showing when suspension phenomena come into the picture of sand movement.

INITIAL MOVEMENT OF SAND GRAINS

So far as sand movement by wind or water is concerned, especially in soil erosion, the most important problem is that of determining the conditions for initial movement. Since it is the pressure and drag forces exerted on the individual grains which start their movement, the problem fundamentally resolves itself into a study of the hydrodynamics of sand grains. Most experimental studies to date have been concerned only with average conditions, that is, with average critical tractive force and average critical velocity. This does not seem to be a fundamentally sound approach, since it appears that the variations in local shear and local velocity due to turbulence are of tremendous significance in the start of sand movement. Bagnold,[1] Shields,[6] and White[8] have analyzed the forces acting on individual grains, and performed experiments determining the constants in their equations for critical tractive force.

Analyses of the equilibrium of sand grains (FIGURE 1) indicate that the critical force on any single grain will be

$$f_c = \alpha \tan \phi \rho' g d^3 \qquad (2),$$
$$\rho' = (\rho_s - \rho),$$

where ρ_s and ρ are densities of sand and fluid respectively, ϕ the angle of repose of the grains, d the particle diameter, g gravity, α a constant depending on the type of flow around the grain. At low Reynolds numbers the force acting on the grain is predominatingly tangential, but at higher numbers the fluid force is predominatingly normal to the grain. This results in a vertical shift in the resultant line of action of the fluid force on the grain.

In uniform sand, if the number of grains per unit area which are effec-

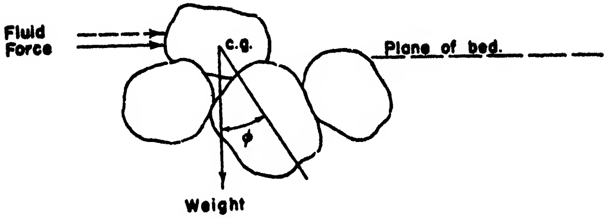


FIGURE 1. Forces on sand grain.

tive in taking the fluid drag is equal to n/d^2, the drag per unit area is equal to $f_c n/d^2$. The value of n depends on the closeness of packing of the grains and can be evaluated by counting the grains which are lying on the top of the bed. If the sand is not uniform in size, an estimate must be made of the proportion of the total drag taken by the sand grains of the size being considered. The critical drag force per unit area is then

$$\tau_c = \alpha n \tan \phi \rho' g d \qquad (3),$$

when the value of $R = U_* d/\nu$ (where $U_* = \sqrt{\tau_0/\rho}$ and ν is the kinematic viscosity) is less than about 3.5. White's experiments indicate that α is about 0.20. When the parameter R is larger than 3.5, the individual grains begin to shed small eddies, which tend to produce a fluctuation in the force acting on the grains. This occurs even though the main stream may be non-turbulent. Thus the critical value of bottom shear will depend on whether the so-called "Reynolds number," R, of the grain is greater or less than 3.5. Experiments to determine τ_c for various materials and conditions can probably be made best in contracting water tunnels where sufficient shear can be developed without the creation of large-scale turbulence in the main stream. The use of viscous liquids other than water may also be desirable: For values of R greater than 3.5, White found experimental values of α of about 0.25.

In any case it is apparent that the critical value of unit shear required to move sand particles can be experimentally determined. The problem now remains as to how this value of τ_c is to be applied in case of natural water courses. First, the important item to keep in mind is that, if even momentarily the value of τ_c is exceeded, movement will result. In normal open-channel flow, turbulence will cause the shear to fluctuate considerably. This fluctuation will undoubtedly be related to the velocity fluctuations. Experimental evidence indicates that the velocity fluctuations are distributed according to the Gaussian normal error law. Thus the standard deviation of the horizontal velocity,

$\sigma = \sqrt{\overline{(U - \bar{U})^2}}$, completely describes the fluctuation. It can be shown that values of $(U - \bar{U})$ greater than 3σ will occur only about 0.3 per cent of the time. Therefore, for practical purposes we can assume that the maximum value of $(U - \bar{U})$ will be about 3σ. Very few measurements of σ near boundaries are available, but there are data that indicate that $\sigma/\bar{U}$, where $\bar{U}$ is the mean velocity at the point considered, may be of the order of 1/3. In that case $U_{max} = 2\bar{U}$ approximately, and if the drag on the particle varies as the square of the velocity, it is apparent that we must expect momentary values of drag equal to 4 times the mean value calculated from surface slope and depth of water flow. In fact, it is possible for the local mean drag to be zero with sand movement taking place due to the turbulence fluctuations with the resultant momentary high drag forces.

It thus appears that experiments on bed sediment movement in laboratory or natural channels involving the measurement of average velocity and drag cannot give fundamental data. Rather, what is necessary is the determination of values of σ and an evaluation from such measurements of how and to what degree the shear fluctuates. For a given τ_c the presence or absence of sand movement depends on whether from a statistical standpoint, in the practical case being considered, the value of τ exceeds τ_c for any significant portion of the time. Hence the mean value of the shear may be of no direct significance.

TOTAL TRANSPORT BY SURFACE CREEP

No attempt will be made to develop a complete and final formula for the rate of surface creep or bed-load movement, because existing data are not of the type that can be used to evaluate the constants in any such equation. Only the general form of what is believed to be the correct type of equation will be presented. Considering any single grain, the velocity at which it will move, U_s, at any instant will be equal to $b(U - U_c)$, where U is the fluid velocity acting on it, and U_c the velocity associated with the critical tractive force, τ_c, and b is a numerical constant. If the number of grains per unit area of the size being considered is n/d^2, the rate of transportation per unit time per unit width of bed is

$$G = U_s \rho_s g n d \tag{4}$$

But the value of U_s varies with time, since U fluctuates, and therefore it is necessary to determine its average value. Experiments indicate that the velocity in turbulent flow fluctuates according to the normal error law (FIGURE 2) so that the frequency function of U is

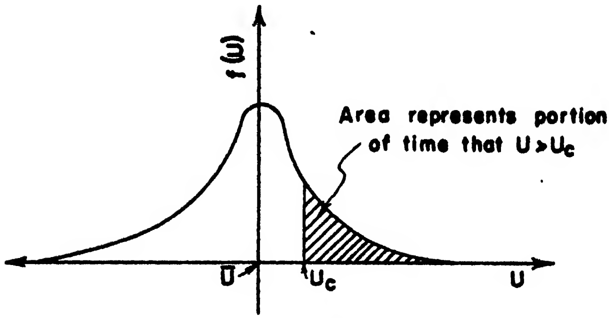


FIGURE 2. The normal error law for fluctuating velocity, U.

$$f(U) = \frac{e^{-(U-\bar{U})^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \tag{5},$$

where $\sigma = \sqrt{\overline{(U - \bar{U})^2}}$, and $\bar{U}$ is the mean fluid velocity at the sand grain level.

The mean value of $\bar{U}_g$ will then be

$$\bar{U}_g = b\int_{U_c}^{\infty}(U - U_c)\,\mathrm{f}\,(U)\,\mathrm{d}U \tag{6}.$$

Let $(U - \bar{U})/\sigma = t$, and $t_c = (U_c - \bar{U})/\sigma$. We then have

$$\bar{U}_g = b(\sigma/\sqrt{2\pi})\int_{t_c}^{\infty} te^{-t^2/2}\,\mathrm{d}t + b(\bar{U} - U_c)\int_{t_c}^{\infty}\frac{e^{-t^2/2}}{\sqrt{2\pi}}\,\mathrm{d}t \tag{7}$$

$$= \frac{b\sigma e^{-t_c^2/2}}{\sqrt{2\pi}} + b(\bar{U} - U_c)\,\mathrm{f}\,(t_c) \tag{8},$$

where $\mathrm{f}(t_c)$ can be found from tables in books on statistics. Note that the value of $\bar{U}_g$ depends on σ, $\bar{U}$ and U_c, and that $\bar{U}_g$ may have a positive value even though U_c is greater than $\bar{U}$.

Where the entire drag is due to the bed, the value of $\sigma/\bar{U}$ tends to be quite constant and certain data on hand indicate that this ratio may be of the order of 1/3. If such is the case, the value of $\bar{U}_g$ as given by equation (8) depends only on U_c and $\bar{U}$; the latter is of course dependent on the grain size and the bottom shear τ_0, and would be evaluated from the velocity distribution expression for flow over rough surfaces.

Perhaps there are data in existence which would permit a checking of equation (4), but the writer has not, as yet, had an opportunity to make a thorough investigation. The physical basis for equation (4) appears fundamentally sound, because it reduces the problem of bed-

load movement to the evaluation of a critical drag or velocity under controlled laboratory conditions, and an evaluation of the turbulence under various natural conditions.

CRITERION FOR SALTATION

As soon as the movement of sand has been initiated, some of the particles may be lifted off the bed into the fluid stream. It can be assumed that they may thus acquire a velocity equal to that of the fluid. An analysis will be made to determine under perfect conditions the maximum height to which a sand particle moving at a specified velocity will bounce after striking the bed. It will be assumed that the particle has acquired kinetic energy equal to $k\rho_s U^2 d^3/2$ and that it will strike a surface such as to cause it to have all its velocity directed straight upward. The height of vertical rise will be rd.

The original kinetic energy will be used in raising the particle against gravity and overcoming fluid friction as its velocity changes from U to 0. The work done against gravity will be $kr\rho' gd^4$. The work done by the fluid resistance in the vertical direction will be equal to the average resistance times rd. The fluid resistance is equal to

$$R = CA\rho U^2/2 \tag{9}$$

where C is a drag coefficient and A the projected area. Let the average resistance be expressed thus:

$$R_{ave} = mCA\rho U^2/2 \tag{10}$$

where m is the ratio between the average resistance and its initial value associated with the initial velocity U. The work done against this resistance will be $R_{ave}\, rd$. On equating the initial kinetic energy to the work done, we have

$$k\rho_s U^2 d^3/2 = rkd^4\rho' g + mCrdA\rho U^2/2 \tag{11}$$

$$r = .5k\rho_s U^2 d^3/[gkd^4\rho' + .5mCdA\rho U^2] \tag{12}$$

Assume that the sand particle approximates a sphere and that the average fluid resistance can be expressed with sufficient accuracy by assuming a linear change in U^2, thus making $m = \frac{1}{2}$. We then have

$$r = 1/[(2gd\rho'/\rho_s U^2) + .75C\rho/\rho_s] \tag{13}$$

Note that the value of r increases with the velocity and decreases with the ratio ρ/ρ_s. In order to compare the value of r for similar conditions for air and water the value of U should be expressed so the shear is identical., Bagnold[1] in his studies of saltation developed an expression for the critical velocity at which sand movement started, viz.

$$U_c = 5.75M\sqrt{\rho' g d/\rho}\,\log\,(30y/e) \tag{14},$$

where M is an empirical coefficient; e is the height of the surface roughness, and U_c the velocity at a vertical distance, y, which causes sand just to begin to move. If the surface roughness is caused merely by the sand grains, $e = d$, and U_c could be calculated for $y = e$. However, for the present analysis let us substitute the value of U_c as given by equation (14) for U in equation (13). The value of r then becomes

$$r = (\rho_s/\rho)/(B + .75C) \tag{15},$$

where $B = .06/M^2 \log^2 (30y/e)$. (Bagnold's value for M was about 0.10.) Thus for conditions of the same value of shear, the value of r depends on the ratio ρ_s/ρ, and therefore, for the same size of particle the height of bounce in water will be only about 1/800 of what it is in air. This apparently indicates that saltation will be many times more significant in air than in water. In fact it does not appear possible, even at high water velocities, for sand particles to bounce up more than a few grain diameters in the saltation process. It may thus be concluded that in water streams, when the velocity reaches sufficient magnitudes to cause saltation, the turbulence will be such as to place the material in suspension and thus entirely obscure any saltation effect.

CRITERION FOR TRANSPORTATION OF SAND AND SILT IN SUSPENSION

Before the wind will transport material in suspension there must be vertical components of velocity fluctuation due to turbulence greater than the velocity of fall of the sand or silt in still air, which will be called c. The analyses for determining the relation between the turbulence, relative concentration at the ground of the sand of size characterized by c, and concentration in suspension at some elevation y, were first made by Lane and Kalinske[4] and later improved by Kirkham.[3] The analysis is made for the equilibrium condition when the rate of picking up of material off the ground by turbulence is equal to the rate of settling. The former must be proportional to the relative amount of such material on the ground, N_b, and the time average of the "picking up" velocity due to turbulence, v. Thus

$$R = A\, N_b \int_c^{\infty} (v - c)\, f(v)\, dv \tag{16}.$$

The average rate of settling is taken as being equal to $N_0 c$ where N_0 is the concentration in suspension of material at the zero suspension level. It is assumed that $f(v)$ is equal to the normal error law. To evaluate

$\sqrt{\bar{v}^2}$, assuming $\tau_0 = uv$, and letting $K = v/u$, we have $\sqrt{\bar{v}^2} = \sqrt{K}\sqrt{\tau_0/\rho}$. Then letting equation (16) equal N_0c, and $t = v/\sqrt{\tau_0/\rho}$ we have

$$\frac{N_0}{N_b} = \frac{A}{t_c\sqrt{2\pi K}}\int_{t_c}^{\infty}(t - t_c)\, e^{-\frac{t^2}{2K}}\, dt \tag{17}$$

Equation (17) indicates that the ratio, $c/\sqrt{\tau_0/\rho} = t_c$ is a function of the concentration ratio, N_0/N_b. It should be recalled that this expression holds only for equilibrium conditions.

Taking some field data for various rivers and canals, it was possible to calculate N_0, N_b, c, and τ_0; this permitted plotting t_c against N_0/N_b. A smooth curve, FIGURE 3, as determined from equation (17), fitted the data quite well if A was equal to 39 and K equal to 0.27.

Since K is equal to the ratio, v/u, it would be of interest to check this with actual determinations of this ratio. At a short distance above the ground in the atmosphere, Taylor[7] found for this ratio 0.30. Some data obtained by the author in pipes indicates the same order of magnitude for K.

TURBULENCE AND DIFFUSION

By analogy with molecular-diffusion theory, it is possible to study the diffusion process in turbulence both analytically and experimentally. If we define by Y the distance traveled above or below the x-axis of a molecular particle in time, t, the following relations hold

$$Y^2 = 2Dt \tag{18}$$

and

$$\frac{dY^2}{dt} = 2D,$$

where D is a diffusion coefficient.

G. I. Taylor[6] derives an analogous expression for movement in a turbulent fluid by introducing a correlation coefficient, r, defined as $v\,v_t/\bar{v}^2$ where v is the velocity of a particle at a certain instant and v_t its velocity a later time, t. Letting $x = \bar{U}t$, where x is a downwind distance and $\bar{U}$ the mean velocity, Taylor gives

$$\frac{dY^2}{dx} = \frac{2\bar{v}^2}{\bar{U}^2}\int_0^x r\,dx \tag{19}$$

When $x = 0$, then r = unity and as x increases, r approaches zero, and thus $\int_0^x r\,dx$ becomes a constant. For large values of x, let the integral be represented by the constant x_0. Equation (19) can then be integrated and we have

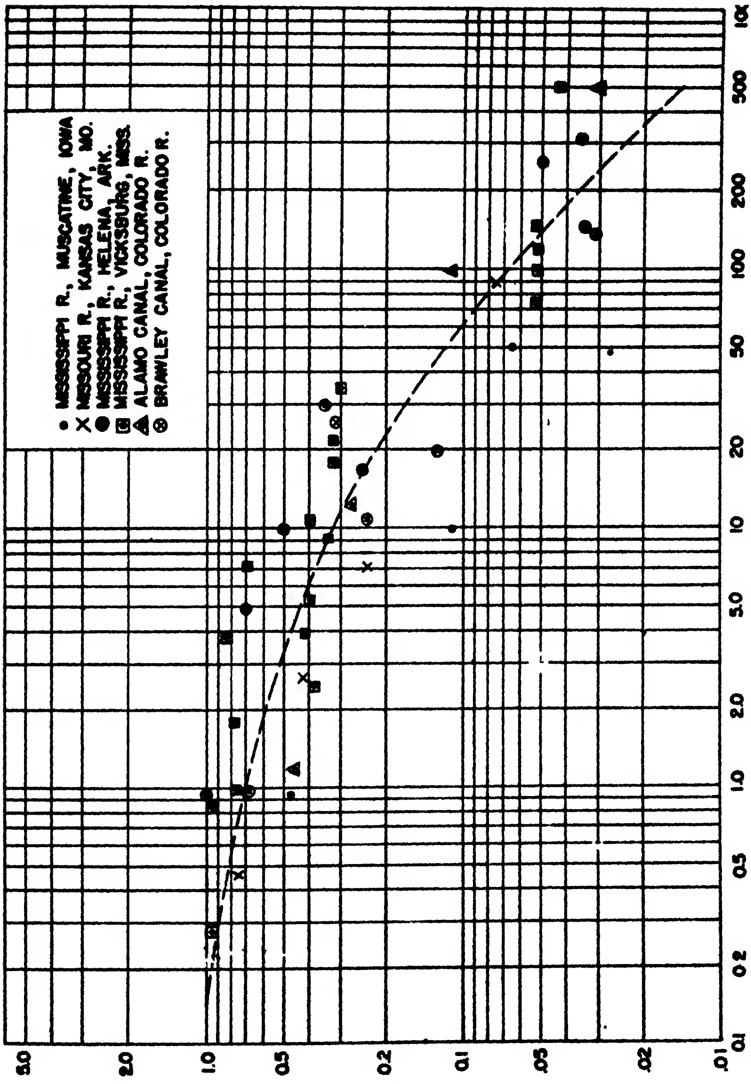


FIGURE 8. Data correlating bed and suspended material with sand characteristics and fluid drag on sand bed.

$$\overline{Y^2} = \frac{2\overline{v^2}x_0}{\overline{U}^2}(x - x_0) \qquad (20).$$

Thus for large values of x, the relation between $\overline{Y^2}$ and x is linear, and the intersection of the straight line with the x-axis gives the constant x_0. This constant is a measure of the scale of the turbulence so far as diffusion processes are concerned. By analogy with molecular diffusion theory, Taylor defined a turbulence diffusion coefficient, D, as being equal to $\overline{U}/2\ (\mathrm{d}\overline{Y^2}/\mathrm{d}x)$, when x is such that $r = 0$. Thus D can be determined from a $\overline{Y^2}$ against x plotting by measuring the slope of the straight line. Note that $D = \overline{v^2}\, x_0/\overline{U}$.

The values of $\overline{Y^2}$ have been experimentally measured in flowing water by injecting immiscible droplets having the same specific gravity as water (a mixture of carbon tetrachloride and benzine has been used), or by injecting a chemical solution having the same density as water (mixture of hydrochloric acid and alcohol used). The droplets were photographed on motion picture film and by projecting the film, the value of $\overline{Y^2}$ could be calculated for various values of x by measuring the values of Y for a large number of particles, and then averaging. Using the chemical solution a slightly different and less time-consuming method was used. Theory and experiment indicate that the mean concentration of the chemical for various values of Y should be

$$C = \frac{M}{\sqrt{2\pi}\,\sqrt{\overline{Y^2}}}\, e^{-Y^2/2\overline{Y^2}} \qquad (21),$$

when $Y = 0$, $C_0 = M/\sqrt{2\pi}\,\sqrt{\overline{Y^2}}$, and therefore $C/C_0 = e^{-Y^2/2\overline{Y^2}}$ or

$$Y^2 = -4.606\, \overline{Y^2} \log C/C_0 \qquad (22).$$

Plotting measured values of C/C_0 against Y^2 permits easy determination of $\overline{Y^2}$. Thus all that was necessary was to obtain several water samples and determine the chloride concentration, which is a relatively simple matter.

In FIGURE 4 are shown data on $\overline{Y^2}$ against x obtained for a variety of conditions of water flow in open channels. Values of D and x_0 are tabulated and the data are plotted in a general dimensionless form. The fact that for any turbulent flow in an open channel the $\overline{Y^2}$ against x curve approaches a straight line clearly indicates that there is a definite limit to the scale of the turbulence. This is in contrast to turbulence in the atmosphere.

In addition to obtaining data on the vertical diffusion coefficient, measurements were also made relating to lateral diffusion. It appeared that near boundaries the lateral diffusion coefficient was consistently

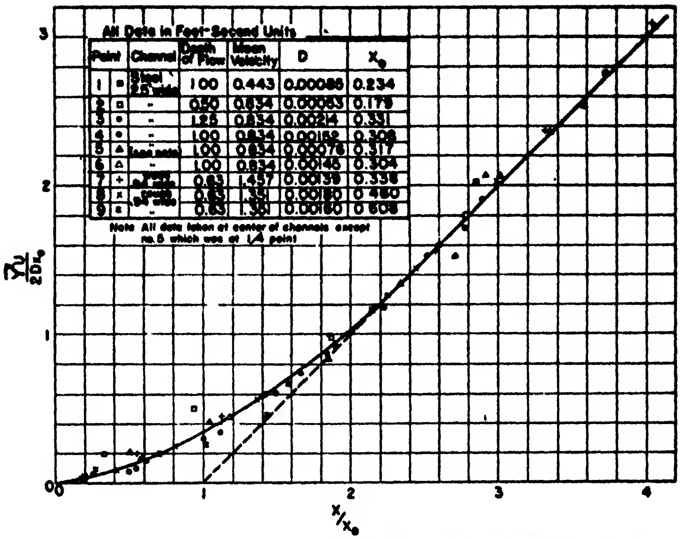


FIGURE 4. Correlation of data on turbulent diffusion for water flowing in open channels.

larger than the vertical coefficient. This is in accordance with data for the atmosphere.

DISTRIBUTION OF SUSPENDED MATERIAL

The suspension problem can be treated as a diffusion phenomenon on which the force of gravity is superimposed. The general equation controlling the phenomenon is

$$\frac{\partial N}{\partial t} = \frac{D_y \partial^2 N}{\partial y^2} + \frac{D_x \partial^2 N}{\partial x^2} + \frac{c \partial N}{\partial y} - \frac{\bar{U} \partial N}{\partial x} \qquad (23),$$

where N is the sediment concentration at a point (x, y) at time t.

The writer has in a previous paper[2] indicated the solution of the above equation for various boundary conditions. The simplest solution is that for equilibrium conditions, in which $\partial N/\partial t$, $\partial N/\partial x$, and $\partial^2 N/\partial x^2$ are zero, and we have

$$cN = D_y \, dN/dy \qquad (24).$$

This expression states that the average rate of upward diffusion of silt by turbulence is equal to the average rate of its dropping due to gravity. The equation integrates to

$$\ln (N/N_a) = - c \int_a^y dy/D_y \qquad (25).$$

The integral can be evaluated if D_y is expressed in terms of y. Experimentally determined values of D_y from previously described diffusion experiments were used in equation (25), and it was found that the distribution curve obtained checked the experimental measurements of sediment concentration quite well. (The detailed results are to be presented in another publication.) In FIGURE 5 are shown typical data on the vertical distribution of mean velocity and vertical diffusion coefficient at the center of a small water channel. In FIGURE 6 are shown plottings of various sizes of sand concentrations in suspension at various levels in this channel. The broken curves are plotted by substituting in equation (25) the values of D_y from FIGURE 5. The reference point for these curves was the concentration at a distance of 0.062 feet above the bottom interpolated from actual values. The correspondence between actual and theoretical variation of the sediment concentration is quite good.

It was found that values of D_y determined in clear water did not correspond to those determined in sediment-laden water. Also, the values of the diffusion coefficient calculated from the velocity gradient and shear relationship did not correspond exactly to the measured values of D_y. This, of course, may be due to the difficulty in accurately

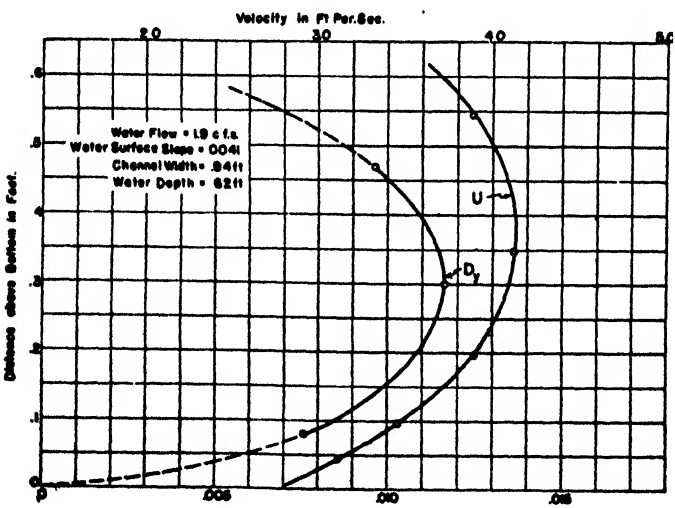


FIGURE 5. Distribution of mean velocity and diffusion coefficient at center of small channel.

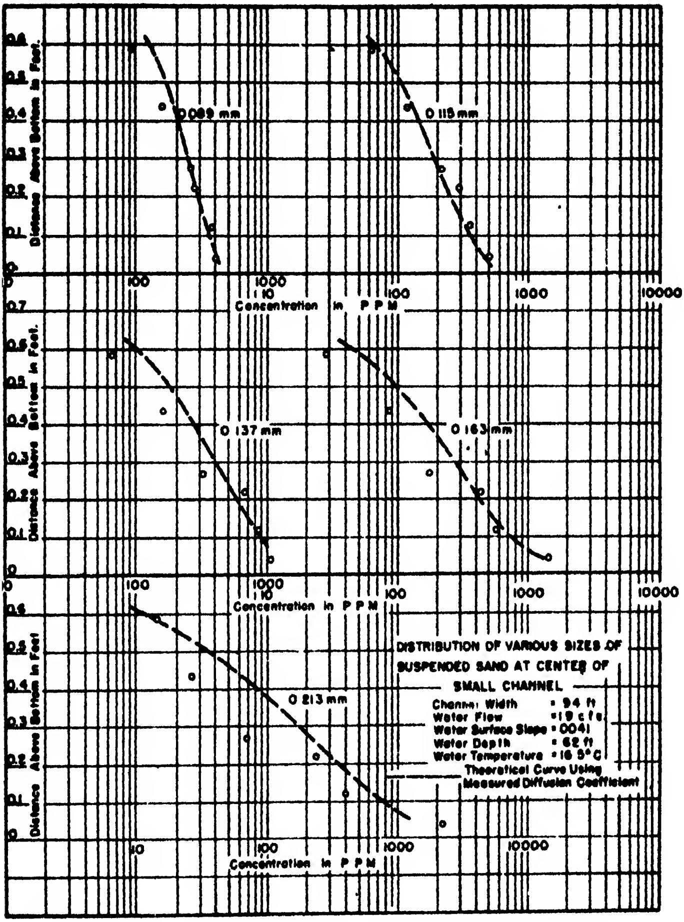


FIGURE 6. Correlation of suspended sand measurements with theory.

calculating values of shear and velocity gradient. In general the measured values of D_y were smaller then the calculated values. It is believed that the direct measurement of both horizontal and vertical diffusion coefficients could be accomplished in the free atmosphere in a manner quite similar to that used in the water channels.

ACKNOWLEDGMENT

The writer wishes to acknowledge the many suggestions made by Professor E. W. Lane regarding the analyses presented herein, and the laboratory assistance of J. M. Robertson and Chung-Ling Pien.

REFERENCES

1. **Bagnold, R. A.**
   1941. Physics of blown sands and desert dunes. London. Methuen & Co.
2. **Kalinske, A. A.**
   1940. Suspended material transportation under non-equilibrium conditions. Trans. Am. Geophys. Un. **1940** (2): 613.
3. **Kirkham, Don**
   1942. Modification of theory on the relation of suspended to bed material in rivers. Trans. Am. Geophys. Un. **1942** (2): 618.
4. **Lane, E. W., & Kalinske, A. A.**
   1939. Relation of suspended to bed material in rivers. Trans. Am. Geophys. Un. **1939** (4): 637.
5. **Shields, A.**
   1936. Application of similarity principles and turbulence research to bed-load movement. Translation on file in Engineering Societies Library, New York, N. Y., of "Mitteilungen der Preussischen Versuchsanstalt fur Wasserbau und Schiffbau." Berlin.
6. **Taylor, G. I.**
   1921. Diffusion by continuous movements. Proc. London Math. Soc. **20**: 196.
7. 1927. Turbulence. Quart. Jour. Roy. Meteor. Soc. **53**: 210.
8. **White, C. M.**
   1940. Equilibrium of grains on bed of streams. Proc. Roy. Soc. London. **174 A**: 322.

BOUNDARY-LAYER PROBLEMS INVOLVED IN SNOW MELT

By Phillip Light
Hydrometeorological Section, U. S. Weather Bureau, Washington, D. C.

The problem of snow melting and its relation to atmospheric conditions is important in several respects. One is the forecasting of run-off from a snow cover during the melting season. Another is the determination of the maximum limit of melting for a given drainage basin in connection with the design of flood-control works. Lastly, from a meteorological standpoint it is desirable to know the influence of a snow cover on an air mass passing over it.

The problem touches on several fields of meteorology and hydrology. The hydrologic phase of the problem deals with the disposal of melt-water, a portion of which may remain in the snow cover while the remainder finds its way into a stream either through channels in the snow, or along the surface of the ground, or by percolation into the ground. The meteorological phase of the problem is to determine the agents responsible for conversion of the snow into water and to relate the rate of conversion to meteorological factors.

Engineers and hydrologists, who have been brought in contact with the practical aspects of melting snow, have contributed most of the literature on the subject in this country and have developed empirical procedures for handling the problem. Since it is obvious that air temperature must have a significant influence on melting, considerable use has been made of a simple temperature relationship known as the degree-day formula, based on a linear relation of melting to temperature in excess of 32° F. But the "constant" in this formula, known as the degree-day constant, derived empirically through laboratory experiments with snow cores, or by correlations of temperature and stream-flow records, is known to vary considerably. Values ranging from 0.01 to 0.20 inches per degree-day have been obtained by various investigators at different times and over different areas. It is obvious, therefore, that other important factors enter into the problem besides air temperature.

FACTORS AFFECTING THE SNOW-COVER ABLATION

It might be of interest to examine at this point the various influences that are brought to bear on the snow cover. Since the snow is an

intermediate layer between the atmosphere and the soil, it is affected by the processes of radiation, conduction, and convection. Heat and moisture may be transmitted to or withdrawn from the atmosphere by turbulent exchange. The snow may be warmed by radiation, both short-wave from sun and sky, and long-wave from cloud layers and water vapor, and it may be cooled by back-radiation. Precipitation, either rain or snow, may convey additional heat. Heat exchange also takes place between the underlying soil and the snow cover.

At first glance it may seem impossible to diagnose all these factors and arrive at usable relationships. Simplications, however, can be made by neglecting trivial factors. Considering the nature of the available data, all we can hope to do is obtain approximate values. We know, for example, that ordinarily the amount of heat required to bring a mass of snow up to the melting point is small compared to the heat required for melting. Except for deep accumulations of snow, therefore, a uniform temperature of 32° F. can be assumed to exist in the snow cover throughout the melting period so that only the heat conducted at the upper and lower surfaces need be considered. Because the heat transmitted from below will be limited by the low conductivity of the soil, our chief interest will lie with the influences of turbulence, radiation, and rainfall at the surface exposed to the air.

Temperature

Since we are concerned with conditions during melting, the temperature of the air in contact with the snow is 32° F. or very close to that value. Above the snow the temperature increases with height up to the limit of the warm-air inversion and then begins decreasing. Rainfall will generally tend to assume the temperature of its surroundings and, in falling through this type of atmosphere, raindrops on reaching the snow surface may be close to the freezing point. Even if it is assumed that the rain temperature is equal to the air temperature at ordinary station-thermometer levels, computations of heat conducted to the snow show only a slight melting effect for moderate rain intensities. For instance, a rainfall of 4 inches during a 24-hour period with 60° F. temperature, which is an extremely high temperature for rain falling over snow cover, will melt only 0.8 inch of water equivalent of snow.

Radiation

Next, we may consider the effect of radiation on melt. Because of selective, absorption by the snow surface, most of the incident short-wave radiation from sun and sky is reflected, but most of the long-wave

radiant energy is absorbed. Snow has a high absorption coefficient so that, except for very shallow snow, a negligible amount of the radiant energy that penetrates the snow reaches the ground surface to be absorbed there. The percentage of radiation reflected at the upper surface, the albedo, depends mainly on the character of the snow surface. For non-melting conditions the albedo is high, but during melting, when liquid water is present at or near the surface, the albedo is considerably less. Olsson made careful observations of the albedo of the snow, and averages of a number of measurements made under different weather conditions show about 75 per cent albedo for frozen snow and about 60 per cent for melting snow. One interesting result is that if melting begins during the day, due to a rise in air temperature, more radiant energy will be absorbed by the snow and the melting rate will increase. In other words, temperature may act as a catalyst to induce melt by radiation.

Although insolational heating of the snow occurs during the period of daylight, the snow cover radiates back to space throughout the 24-hour period, so that the net daily radiation transfer of heat in middle latitudes may be small. At the comparatively low temperature of the snow surface, the snow may be said to act as a black body and to transmit heat at a rate governed by the Stefan Boltzmann law. A portion of this heat is re-radiated by clouds and water vapor so that the net long-wave radiation outward is considerably reduced. Therefore, both short- and long-wave radiation transfers are redu d by virtue of cloud cover and humidity. Computations of radiation can be made by various formulas and graphical procedures. Ångstrom[1] has a formula that relates insolation transmitted through the atmosphere during the day to the number of hours of sunshine, and both Ångström and Brunt[2] give formulas for computing net long-wave radiation in terms of vapor pressure at the surface. These formulas are empirical, and are based on the means of a large number of observations with fairly large deviations from the mean. Elsasser's method of computing graphically long-wave radiation transfer through the atmosphere has the advantage that the vertical distribution of temperature and water vapor may be used. Either of these methods may be utilized to obtain adequate estimates of radiation transfer if observations of temperature, humidity and cloudiness are available.

During normal conditions, in middle latitudes and with clear sky, melting through radiation will occur during the day with refreezing of the top layer of melt-water at night. For abnormal conditions as during storm situations, cloudiness and increased water-vapor content

of the air reduce short-wave radiation transfer, although there may be a question about the magnitude of long-wave radiation transfer from very moist warm air to the snow. But I think it may be safely stated that, in situations where the advection of warm air takes place at a rapid rate over a snow field, the transfer of heat by turbulence will predominate to a great extent; furthermore, considering the fact that usually only rough results are desired because of the errors due to inadequate data, radiation effects may be disregarded during such situations.

Austausch

In the brief discussion of the temperature distribution in the air, it was indicated that the surface temperature with melting snow cover remains constant at 32° F. Since the air immediately above the snow is saturated with water vapor, the vapor pressure of the air at the surface is equal to the saturated vapor pressure of air at 32° F., or 6.11 mb. This simplifies the problem of determining heat and water-vapor exchange, because the surface conditions are known, and indicates the possibility of computing quantities of heat flow and water-vapor flow to the snow surface from a single set of observations in the air layer.

In general terms, if potential temperature increases with height, heat flows downward from air to snow. Similarly, the direction of water-vapor transport is governed by the vapor-pressure gradient. If the vapor pressure at a short distance above the surface is greater than 6.11 mb., or the dew-point is greater than 32° F., water vapor is brought down to the snow surface, where it is condensed and releases latent heat of condensation. Conversely, if the dew-point is less than 32° F., moisture is evaporated into the air and the snow cover loses heat of evaporation. Expressed in quantitative terms, heat transfer is proportional to the product of the *Austausch* of heat, or eddy conductivity, and the potential temperature gradient. For short vertical distances in an inversion, ordinary temperature gradient can be substituted for potential temperature gradient. Water-vapor transfer is proportional to the product of eddy conductivity and vapor-pressure gradient. The problem, then, is to determine the relation between these quantities and the meteorological elements as they are measured at the instrument levels.

For the solution of this problem there are available the researches of Rossby,[9] Sverdrup,[10] and others. Rossby, utilizing the theory of the mixing length due to Prandtl, has shown that for an adiabatic atmosphere, the *Austausch* of momentum, or eddy viscosity, is a linear function both of height, and of wind velocity at a fixed height, through-

out the boundary layer. In accordance with this theory, surface roughness is included in a factor termed the roughness parameter which is assumed to remain constant for all ranges of wind velocity. With a steady vertical flow of momentum or constant frictional stress, the wind-velocity gradient is inversely proportional to the eddy viscosity, and therefore inversely proportional to height. From this it follows that wind velocity must be distributed logarithmically with height and the frictional stress can be related to wind velocity, height above the surface, and the roughness parameter corresponding to the type of surface. The roughness parameter can be determined by plotting observations of wind at several levels on semi-logarithmic paper and extrapolating the wind velocity profile on a straight line to the zero level, where the corresponding height represents the roughness parameter.

Sverdrup has extended this development to the case of heat and water-vapor transfer to melting snow. By assuming a steady state of flow he has derived the following formulas based on logarithmic distributions of temperature and vapor pressure, with the height intercept of surface temperature and vapor pressure equal to the roughness parameter of the snow surface.

$$Q = c_p K U (T - T_0) \qquad (1),$$

$$F = \frac{0.622}{p} K U (e - e_0) \qquad (2),$$

where

$$K = \frac{\rho k_0^2}{\ln \frac{a}{z_0} \ln \frac{b}{z_0}} \qquad (3),$$

where Q = heat exchange, F = water-vapor exchange, c_p = specific heat of air at constant pressure, ρ = density of air, k_0 = von Kármán's coefficient, U = wind velocity, T = air temperature, T_0 = snow-surface temperature, e = vapor pressure of air, e_0 = vapor tension of snow surface in mb., p = atmospheric pressure in mb., a = elevation of anemometer, b = elevation of hygrothermograph, and z_0 = roughness parameter. All quantities not otherwise designated are in c.g.s. units.

For application to a level surface, these formulas require observations at only one level since the roughness parameter of a smooth snow surface, which has been determined by experiment to be 0.25 cm., is assumed to remain constant under all conditions. Sverdrup has, however, suggested that the distribution of elements can be better represented by a power function under conditions of stability with warm air over a cold snow surface. The corresponding formulas necessitate

observations at two or more levels above the snow because the exponent in the formula varies with the strength of the wind, amount of stability, and the height of observations, and must be determined through measurements of the gradients of temperature and wind velocity.

THE ABLATION FORMULA

Since an expression is desired that relates melting to the ordinary type of observations—that is, observations at one level above the ground—the logarithmic type of formula should be checked by actual data. One source of data is the published results of the Norwegian-Swedish Expedition of 1934 to Spitzbergen, where measurements of ablation of snow accumulated on glaciers were made simultaneously with meteorological observations at several levels. Measurements of sun and sky radiation and the albedo of the snow were also made so that the portion of the observed ablation contributed by atmospheric turbulence can be identified. Corresponding values of ablation from meteorological observations can be computed through the use of the following relation, expressing the total heat conducted by sensible-heat transfer and water-vapor transfer in terms of depth of snow ablated at the surface,

$$D = \frac{Q + 600F}{80} \qquad (4),$$

where D is ablation in cm./sec.

Substituting for surface temperature and vapor pressure in equations (1) and (2), and then substituting for Q and F in the above equation, we obtain

$$D = \frac{KU}{80}\left[c_pT + \frac{373}{p}(e - 6.11)\right] \qquad (5).$$

Equation (5) was used to compute two sets of ablation values, one from observations of wind at 7 meters and temperature and humidity at 5 meters, and the other from observations at 2 meters and 1 meter, respectively. The results, in the form of equivalent heat units of computed ablation plotted against observed ablation, are shown in FIGURE 1. It can be seen that the points fall on either side of the line of perfect agreement and no systematic error can be detected from the limited amount of data. The chart does show a pronounced tendency for values of high wind speeds, namely, points 1, 2, and 10, to agree more closely. This seems to show that the logarithmic formula tends toward greater accuracy with increased wind speed.

Theory indicates that the greater the wind speed the more nearly

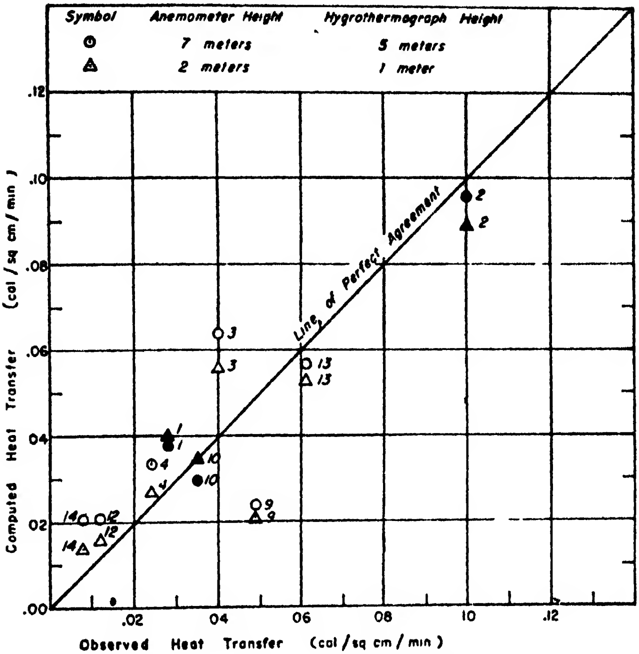


FIGURE 1. Computed heat transfer vs observed heat transfer for various melting periods

eddy conductivity approaches a linear relation with height, and therefore, the greater the accuracy of the logarithmic formula. On the other hand, the greater the temperature of the air the greater is the stabilizing force that opposes turbulence and so the formula should become less accurate when applied to higher temperatures. These facts indicate limitations to the indiscriminate use of the formula for all ranges of wind speed and temperature.

For practical use in determining runoff, equation (5) should be modified to include water condensed or evaporated. The resultant formula then becomes

$$D = \frac{Q + 600F}{80} + F = \frac{KU}{80}\left[c_pT + \frac{423}{p}(e - 6.11)\right] \qquad (6),$$

and here D may be termed effective snow melt and represents the net contribution to runoff from turbulent exchange.

Further simplifications can be made by adopting reference levels of instruments of 50 ft. for a, and 10 ft. for b, and evaluating K by the following substitutions: $c_p = 0.24$, $k_0 = 0.38$, $z_0 = 0.25$. Also, the following approximate relation between pressure or density and elevation, applicable between sea-level and 10,000 feet, may be used.

$$\frac{p}{p_0} = \frac{\rho_0}{\rho} = 10^{-.016h} \tag{7}$$

where p_0, and ρ_0 are normal sea-level pressure and density, and h is elevation above sea-level in thousands of feet. The resultant formula expressed in convenient units becomes

$$D = .0018U\left[(T - 32)\,10^{-.016h} + 3.2\,(e - 6.11)\right] \tag{8}$$

where D is effective melt in inches per six hours, U is average wind velocity in miles per hour, T is temperature in °F., and e is vapor pressure in millibars. Ordinarily, in view of the approximations involved in the formula, it is unnecessary to correct for instrument elevations. However, if the actual levels of instruments differ considerably from the normal elevations selected as reference levels in deriving the formula, the right hand side of equation (8) is multiplied by the correction term $\dfrac{11.7}{\log_{10}(1.22a)\log_{10}(1.22b)}$, where a and b are in units of feet.

For convenience of computations, a graph has been constructed applicable to lowland drainage basins (FIGURE 2) giving values of effective snow melt in terms of temperature and relative humidity for a unit wind velocity of one mile per hour. Since the formula gives melting rates as a linear function of wind velocity, values read from the graph are multiplied by the observed wind velocity.

ABLATION FROM A DRAINAGE BASIN

The previous discussion has dealt entirely with melting over a limited area near the location of the observation station. In practice, however, we are concerned with melting rates over a fairly large area extending over several hundreds or even thousands of square miles. The snow surface is not uniformly smooth, the entire area is not at the same elevation as the meteorological station, and some of the basin area may be sheltered from the direct influence of turbulence by forests or by ridges. Hence considerable modifications are necessary in dealing with snow melt over large drainage areas. Elevation differences offer no great difficulty because they can be accounted for by assuming a constant equivalent-potential temperature and a constant pressure gradient over

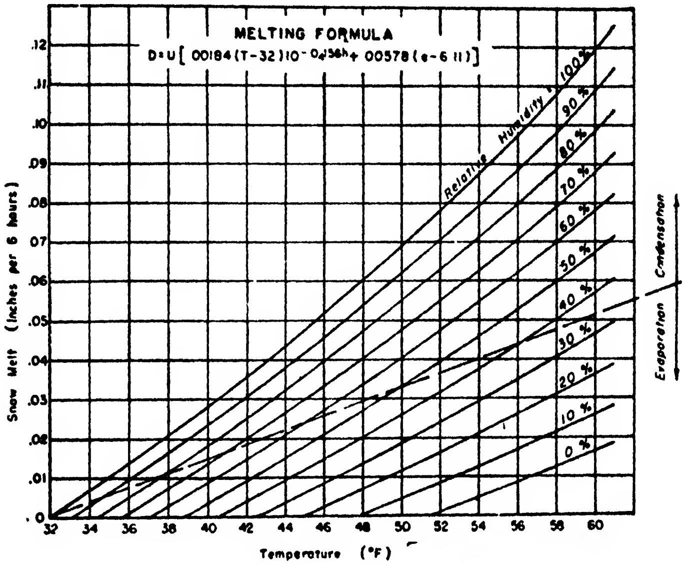


FIGURE 2. Effective snow melt due to turbulent exchange for unit wind velocity.

the basin, and then correcting observations to the mean elevation of the basin.

A question that offers greater difficulty is this: If the area were perfectly level can one assume a uniform condition of the air mass over the basin? It is obvious that if the air is transmitting heat and moisture to the snow, the air in turn is losing heat and moisture and consequently should suffer a reduction of temperature and humidity along its path. This means that unless the observation station is centrally located its readings are not a true index to meteorological conditions. A station is usually located in the valley of the basin so that it normally measures conditions of the air as it enters or leaves the basin. Hence, if a considerable reduction of temperature and vapor pressure did take place over a snow field, our predicted values of snow melt would be in error.

This question cannot be answered at the present time with any degree of assurance. One of the bothersome points is the extent to which the heat and moisture transfer takes place in the air; that is, the thickness of the layer that participates in the transfer. The author has computed rates of reduction of melting power of the air by a method described in a preceding paper. It utilizes Rossby and Montgomery's

formula[8] for the height of the turbulent layer in an adiabatic atmosphere, and assumes that the heat loss and water-vapor gain or loss along the trajectory of the air is distributed uniformly throughout the turbulent layer. But this method has obvious defects because the layer of air is stable and the change in properties probably is not distributed uniformly. A mathematical treatment of this problem should take into account the variation of eddy conductivity with height in the whole frictional layer. The only thing that can be said, at present, is that it appears from the limited data available that the effect is small for high wind velocities. There are records of high air temperatures above deep snow with no marked increase of temperatures in the vicinity above bare ground.

Questions can be raised regarding the influence of the surface characteristics of a drainage basin. We can distinguish two such factors that would affect the results from a melting formula based on turbulent transport: surface roughness and forest cover. Since the snow surface tends to follow the contours of the ground, although a certain amount of smoothing would certainly occur, the original character of the basin surface would have to be considered in any theoretical analysis of the problem. In the logarithmic type of formula, however, roughness enters as a constant factor so that the air properties do not affect the particular roughness parameter of the surface. This indicates that a previously determined value of an over-all roughness parameter for the watershed could be employed as a basin constant and enter into the formula as a simple multiplier. Since we can only hope to evaluate this constant empirically, the evaluation would have to be from past records of melting used in conjunction with meteorological observations.

Forest cover introduces another complication, and the effect of type and relative area covered by forest in a basin should be considered. We know that wind speed is reduced in the interior of a forest to a degree dependent on the height and density of trees and foliage. Geiger[4] has found that the wind in the interior of an open forest has an approximately constant ratio to the wind above the tree tops. Turbulence is considerably reduced in a dense forest, possibly to a point where turbulent transfer may be only a minor influence on heat exchange between the air inside the forest layer and the snow surface. Very little investigation has been made of eddy conductivity inside a forest, although work has been done on the distribution of temperature, wind velocity and humidity. These investigations indicate that the forest reduces the heat exchange to a fraction of that over an open area. Applying this idea to the problem of melting underneath a forest, the following observations might be made. In a dense forest, heat exchange

is reduced to the point where the snow melts at an insignificant rate. In an open forest, where turbulence of the air is an important factor, the rate of melting is governed by the turbulence regime and will be proportional to melting in a nearby exposed area.

All these considerations add up to the assumption of an over-all empirical factor prefixing the formula, constant for a particular basin or homogeneous region. The Hydrometeorological Section of the U. S. Weather Bureau has made several investigations of drainage-area melt in different sections of the country and has found this constant to be fairly conservative. Values have been obtained that range from 0.5 to 1.0. The greatest difficulty is to find cases for which constants can be evaluated. Certain conditions are necessary, such as a period of rapid melt in which both temperatures and wind velocities are high. There should also be plentiful data on stream flow, snow depth within the drainage basin, and meteorological elements. In practice, very few test cases that fulfill these requirements are discovered in any particular area of study, so only a limited number have been analyzed as yet. The validity of the method is indicated by the fact that the distribution of melt computed by means of the formula, and the introduction of a basin constant, conforms reasonably to the actual distribution of snow-melt runoff as determined from stream-flow records. Here it should be pointed out that the question of snow-melt runoff involves other problems that may reasonably be expected to obscure the effect of melting at the surface of the snow. The amount of liquid water that might be contained in the snow blanket has been given little attention, but it certainly might have a pronounced effect on the way that melt water makes its appearance in the stream channels. Studies made thus far seem to indicate that melting has the major effect on runoff, and that the runoff characteristics of the basin, as well as the retaining effect of the snow cover itself, are of minor consequence.

An illustration of an actual case of the application of the melting formula and the procedure discussed here is shown in FIGURE 3. Values of computed melt for the period, March 21-28, 1936, in the French Creek Basin at Saegerstown, Pennsylvania, were accumulated daily from the time melting began and are plotted on a time scale. Alongside, on the same scale are plotted values of snow-melt runoff obtained from runoff analysis. It can be seen that, except for a time lag between the two curves, the shapes of the curves agree. Here a basin constant of 0.65 was used, obtained by the analysis of a number of basins in the same region, the upper Ohio watershed.

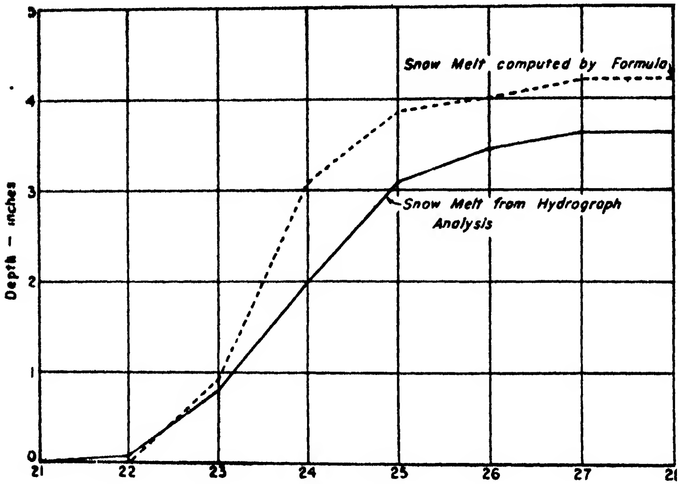


FIGURE 3. Mass curves of snow melt for French Creek Basin (620 sq. miles) at Saegerstown, Pa., March 21–28, 1936.

SUMMARY

Summarizing the results of investigations made thus far, we may say that although the application of the procedure described here shows promise, numerous problems remain, both from a theoretical standpoint for an ideally flat snow surface and from a practical standpoint for areal melt over actual drainage basins. One of them is the determination of the errors involved in the use of a simple melting formula of this type. For instance, is melting a linear function of wind velocity for all conditions of wind and temperature? Also, how does the area of the snowfield affect this linear relation? Sutton's investigations of evaporation over plane areas indicate that velocity to some exponent less than one is a truer proportionality for eddy transport from surfaces of various shapes and areas. Other aspects of melting are worth consideration. The effect of rainfall on melting has been dealt with briefly and it was indicated that a complicating factor is the unknown value of rain temperature. Also, long-wave radiation transfer from very moist air close to the snow surface needs further study.

The great need in furthering a study of this kind is some sort of experimental setup. Perhaps an experimental basin can be found in

which measurements can be made of turbulent transport, outgoing and incoming radiation, ablation of snow, and stream discharge. Such an area set aside for a period of time for an experimental investigation would provide valuable data for settling some of the problems outlined here.

ACKNOWLEDGMENT

The writer gratefully acknowledges helpful assistance rendered by Dr. H. Wexler, A. K. Showalter, and members of the Hydrometeorological Section of the U. S. Weather Bureau.

DISCUSSION OF THE PAPER

Dr. R. B. Montgomery (*New York University, New York, N. Y.*):

Sverdrup's study of conditions over the snow field in Spitzbergen indicated that the effect of stability on eddy diffusivity, which may be very great at the elevation of common anemometer exposures, is negligible at elevations less than about 1 meter above the snow surface. In studies such as Mr. Light's, therefore, it appears highly desirable that the instruments for measuring temperatures and wind be located close to the surface, probably between 1 and 2 meters above the snow. Under stable conditions the logarithmic distribution of wind, temperature, and humidity can be assumed for only this short distance above the surface.

Reply by Mr. Light:

Ordinary instrument elevations at Weather Bureau stations are greater than two meters above the ground, so that a formula restricted to observations below that level would not be of much practical use. However, FIGURE 1 does not indicate large discrepancies between computations of heat transfer made by means of the logarithmic formula from observations of wind at seven meters and temperature at five meters as contrasted to computations using data at corresponding heights of two meters and one meter.

Mr. Allan C. Clark (*Pan American Airways, La Guardia Field, N. Y.*):

As an aid to discussion of the French Creek Valley experiment, the following is offered, based on personal acquaintance with the area.

The valley through which French Creek runs is narrow in many places, with hills and forests on either side. Open ground occurs mostly in small areas. Snow fall is locally variable, often decreasing from north to south as a result of decreasing effect of "instability" snow showers with increasing distance from the shores of Lake Erie. These factors might unduly complicate the use of snow runoff data from this area as a test of usefulness of snow-melt formulas.

Dr. H. Wexler (*U. S. Weather Bureau, Washington, D. C.*):

When snow cover is not too thick (< 12 in. depending on the intensity of insolation and properties of the snow) the little insolation that does penetrate the snow is completely absorbed by a top skin-layer of the ground. This may raise its temperature to 32° F. and thus begin melting of the snow from below. This may result in formation of "snow pockets" (depressions in the snow cover) in regions where insolation is large.

Reply by Mr. Light:

Sunlight that penetrates through shallow snow and is absorbed at the ground surface does not increase the melting *rate*. It means only that some of the melting will occur at the lower surface.

Prof. C. F. Brooks (*Blue Hill Observatory, Milton, Mass.*):

I note that several factors have been neglected, as perhaps usually insignificant: (1) the heat required to bring the snow to the melting point; (2) the heat required to bring the surface of the ground to the freezing point and to keep it there (offsetting conduction); (3) the amount of liquid water in the snow cover—up to 30 per cent in fairly fresh melting snow, according to calorimetric measurements by C. F. Merriam (Trans. Am. Geophys. U., 1941) and others, including myself; (4) the relatively difficult penetration of the wind to the smooth cold snow surface, becoming increasingly difficult the warmer the wind is, as indicated by observed great contrasts in air temperature close to the snow surface and only a few feet above it. Items (1), (2), and (4), above, may help explain why computed melt in March 1936 exceeded the observed.

Reply by Mr. Light:

The quantity of negative heat storage in snow is small (unless an intense prolonged period of cooling precedes the melting period), because of the low rate of heat diffusion through the snow. Liquid water in the snow includes water held by capillary action and gravity water. The latter is in process of draining off from the snow cover and should be considered as temporary storage. Latent-heat measurements, therefore, do not indicate net amount of water held in permanent storage by the snow.

REFERENCES

1. **Ångström, Anders**
   1918. On the radiation and temperature of snow and the convection of the air at its surface. Arkiv f. Matematik, Astronomi och Fysik **13** (21): 17-18.
2. **Brunt, David**
   1939. Physical and dynamical meteorology. Cambridge University Press.
3. **Clyde, G. D.**
   1931. Snow-melting characteristics. Utah Agric. Exp. Stat. Bull. **231**.
4. **Geiger, Rudolph**
   1927. Das Klima der Bodennähe Luftschicht. Die Wissenschaft **78**.
5. **Horton, R. E.**
   1915. The melting of snow. Monthly Weather Review **43** (12): 599.
6. **Light, Phillip**
   1941. Analysis of high rates of melting. Trans. Am. Geophys. Un. **1941** (1): 195-205.
7. **Olsson, Hilding**
   1936. Radiation measurements on Isachsen's Plateau. Geografiska Annaler **18**: 225-244.
8. **Rossby, C.-G., & Montgomery, R. B.**
   1935. The layer of frictional influence in wind and ocean currents. Mass. Inst. Tech. Meteor. Papers **3** (3).
9. **Sutton, O. G.**
   1934. Wind structure and evaporation in a turbulent atmosphere. Proc. Roy. Soc., London. **146 A**.
10. **Sverdrup, H. U.**
    1936. The eddy conductivity of the air over a smooth snow field. Geofysiske Publikasjoner **11** (7).

THE EFFECT OF A GRADUAL WIND CHANGE ON THE STABILITY OF WAVES

BY B. HAURWITZ

Department of Meteorology, Massachusetts Institute of Technology, Cambridge, Massachusetts

The investigation of stability and instability of waves is the main problem of the wave theory of cyclones. In studying the possibility of unstable waves in the atmosphere it has, as a rule, been assumed that a sharp surface of discontinuity separates two fluid masses of different density and velocity. In reality, however, the change from the values in the one layer to those in the other layer is never abrupt, owing to the effects of viscosity, molecular or turbulent. A general discussion of the effects of viscosity on the wave motion and on its stability leads to a rather complicated problem which is beyond the scope of this paper. Instead, only the effect of viscosity in the transitional layer will be discussed. Since the most important cause of unstable waves is a velocity shear, gravitational and dynamic instability being much less important, only the shearing instability will be considered.

SUMMARY

When a sharp discontinuity of the velocity exists, shearing waves are always unstable in the absence of a stabilizing influence, such as a density discontinuity. But if the velocity changes gradually through a transitional layer, the wave motion is unstable only if the wave length is at least five times as large as the width of the zone of transition as shown in equation (20). This result, which already had been derived by Lord Rayleigh[5] in a discussion of the sensitivity of jets to sound, will be shown to hold also when the effect of the earth's rotation is taken into account. Consequently, the unstable region of wave lengths favorable for cyclone formation is narrowed, and more so if the frontal zone is wide, a result which might be expected *a priori*.

The phase difference between the components of the velocity of the wave motion normal to the front is 90° in the case of unstable waves at a sharp discontinuity. If the two layers are separated by a zone of transition the phase difference between the two main layers is somewhat larger than 90° and the variation of the phase is not abrupt but continuous through the zone of transition.

At least in temperate latitudes the wind velocity increases, as a rule, up to the tropopause, but in the stratosphere it decreases again. At the tropopause there is no sudden change of the wind itself but only of the vertical wind variation. Such a wind distribution does not give rise to shearing instability as shown in equation (24). Under these circumstances waves will not form spontaneously at the tropopause but will rather be of a secondary nature, e. g., produced by waves at the frontal surface in the troposphere.

ZONE OF TRANSITION BETWEEN TWO LAYERS OF UNIFORM MOTION

To study the behavior of the shearing instability when the wind shear is replaced by a finite zone of transition in which the wind changes steadily from the value in one to that in the other layer, the following fluid system will be considered (FIGURE 1). In the first layer, which

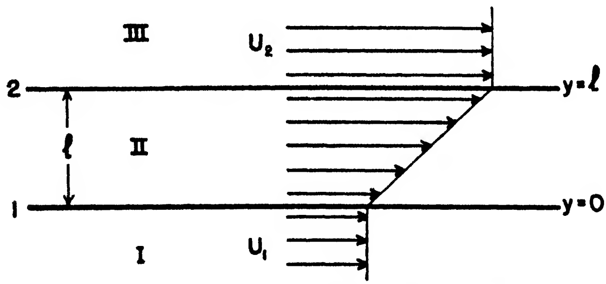


FIGURE 1. Transition layer (II) between two layers, (I) and (III), of different velocity (Arrows proportional to velocity.)

extends from $y = -\infty$ to $y = 0$, the wind velocity is uniformly U_1. In the second layer, which extends from $y = 0$ to $y = l$, the wind speed changes gradually from U_1 at $y = 0$, to U_2 at $y = l$. In the third layer, extending from $y = l$ to $y = \infty$, the wind velocity is uniformly U_2. The density is assumed to be the same in all three layers, as only the shearing instability will be considered.

It will be assumed further that the wave motion occurs in a horizontal plane. To simplify matters, only the vertical component of the earth's rotation is taken into account, a procedure which is strictly correct only at the earth's poles.

The two main layers, I and III, are regarded as infinitely wide in order to simplify the discussion. Moreover, finite layers do not have

essentially different stability conditions. The following two relations hold for the undisturbed motion in each layer:

$$fU = -\frac{1}{\rho}\frac{\partial P}{\partial y} \qquad (1),$$

and

$$g = -\frac{1}{\rho}\frac{\partial P}{\partial z} \qquad (2),$$

where $f = 2\omega \sin \phi$ is twice the vertical component of the angular velocity of the earth's rotation, g is the acceleration of gravity, ρ is the density, and P and U are the pressure and velocity of the undisturbed motion. The rate of shear throughout the transitional layer II may be constant,

$$U = U_1 + by, \qquad \text{when } 0 < y < l \qquad (3),$$

and

$$b = \frac{U_2 - U_1}{l} \qquad (3a).$$

Such a linear velocity distribution represents quite well the actual velocity distribution established in the transitional layer by viscosity, except in the neighborhood of the boundaries, where the present assumption leads to discontinuities of the wind shear which are, however, without great importance in a discussion of the stability. For the equilibrium pressure in layer II the following expression is found from equations (1), (2) and (3):

$$P = P_0 - f\rho\left(U_1 y + \frac{b}{2}y^2\right) - g\rho z \qquad (4).$$

To obtain P for layer I, $b = 0$ has to be substituted in this equation; to obtain P for layer III, U_1 must be replaced by U_2. The free upper surface of the fluid system is given by the condition that P is a constant. Denoting the height of the free surface by h, it follows for layer II that

$$h = H - \frac{f}{g}\left(U_1 y + \frac{b}{2}y^2\right) \qquad (5).$$

The constant H represents the height of the free surface where $y = 0$. Corresponding expressions hold for layers I and III.

If u, v and p stand for the components of the perturbation velocity and the perturbation pressure, the equations of motion and of continuity, neglecting terms of the second order, become,

$$\frac{\partial u}{\partial t} + U\frac{\partial u}{\partial x} + bv - fv = -\frac{1}{\rho}\frac{\partial p}{\partial x} \qquad (6),$$

$$\frac{\partial v}{\partial t} + U\frac{\partial v}{\partial x} \qquad + fu = -\frac{1}{\rho}\frac{\partial p}{\partial y}$$

and

$$\frac{\partial(hu)}{\partial x} + \frac{\partial(hv)}{\partial y} = 0.$$

In order to apply these equations to the three fluid layers, proper values have to be assigned to the parameters U and b for each layer. The equation of continuity can be written in view of equation (5)

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} - \frac{lU}{gh}v = 0 \tag{7a}$$

When the x-axis is chosen in the direction of the wave propagation,

$$\mu, v, p \propto e^{i(\mu x - \nu t)} \tag{8}$$

where $\mu = \frac{2\pi}{L}$, $\nu = \frac{2\pi}{\tau}$, L and τ being the wave length and the period respectively, and $i = \sqrt{-1}$. Then

$$\frac{2\pi}{L}.$$

The third term in equation (7a) can be neglected to a sufficient degree of approximation if L is not too large. For, if $L = 2000$ km,

$$\frac{\partial}{\partial x}\Big| \sim 3.1\ 10^{-6}\, m^{-1}.$$

On the other hand, the factor $\frac{fU}{gh} \sim 1.25\ 10^{-8}\, m^{-1}$, if $U = 10$ m/sec and h is equal to the height of the homogeneous atmosphere. Since u and v, and $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ are of the same order of magnitude, the third term is considerably smaller than the other two, and equation (7a) may be simplified:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{7b}$$

From this and equation (8) it follows that

$$\nu = \frac{i}{\mu}\frac{dv}{dy} \tag{9}$$

According to the first equation (6) and equation (9)

$$\frac{p}{\rho} = i\frac{\nu - \mu U}{\mu^2}\frac{dv}{dy} + \frac{i}{\mu}(b - f)\,v \tag{10}$$

Substituting equations (9) and (10) in the second equation (6), an equation for v is obtained

$$\frac{d^2v}{dy^2} - \mu^2 v = 0 \qquad (11).^*$$

The parameter f which represents the effect of the earth's rotation, does not appear in this equation but only in the expression (10) for the pressure. This is due to the fact that the simple continuity condition (7b) holds with a sufficient degree of approximation. For, if the variation of the vorticity is formed from the equations (6) by cross differentiation, the term $f\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)$ appears.

The following boundary conditions are to be satisfied. The solution for layer I must not become infinite when $y \rightarrow -\infty$, and the solution for layer III must not become infinite when $y \rightarrow \infty$. Hence

$$v^{\mathrm{I}} = K^{\mathrm{I}} e^{\mu y},\ \frac{p^{\mathrm{I}}}{\rho} = i\left(\frac{\nu - \mu U_1}{\mu} - \frac{f}{\mu}\right) K^{\mathrm{I}} e^{\mu y} \qquad (12);$$

$$v^{\mathrm{III}} = K^{\mathrm{III}} e^{-\mu y},\ \frac{p^{\mathrm{III}}}{\rho} = -i\left(\frac{\nu - \mu U_2}{\mu} + \frac{f}{\mu}\right) K^{\mathrm{III}} e^{-\mu y} \qquad (13);$$

while for the transitional layer

$$v^{\mathrm{II}} = K_1^{\mathrm{II}} e^{\mu y} + K_2^{\mathrm{II}} e^{-\mu y}$$

$$\frac{p^{\mathrm{II}}}{\rho} = i\,\frac{\nu - \mu U^{\mathrm{II}}}{\mu}\left(K_1^{\mathrm{II}} e^{\mu y} - K_2^{\mathrm{II}} e^{-\mu y}\right) + \frac{i}{\mu}(b - f)\left(K_1^{\mathrm{II}} e^{\mu y} + K_2^{\mathrm{II}} e^{-\mu y}\right) \qquad (14).$$

Here U^{II} stands for the wind distribution in the transitional layer. K^{I}, K^{III}, K_1^{II}, K_2^{II} are constants which are related by the boundary conditions at the boundaries of layer II. Since there are neither sharp discontinuities of the wind nor of the density separating this layer from the other two layers, the y-components of the velocity and the pressures must be continuous at both boundaries.

$$v_1^{\mathrm{I}} = v_1^{\mathrm{II}} \qquad v_2^{\mathrm{II}} = v_2^{\mathrm{III}}$$

$$p_1^{\mathrm{I}} = p_1^{\mathrm{II}} \qquad p_2^{\mathrm{II}} = p_2^{\mathrm{III}}$$

The upper indices refer to the layer, the lower indices to the boundary. From the condition that the velocities must be continuous, it follows that

$$K^{\mathrm{I}} = K_1^{\mathrm{II}} + K_2^{\mathrm{II}},$$

$$K^{\mathrm{III}} e^{-\mu l} = K_1^{\mathrm{II}} e^{\mu l} + K_2^{\mathrm{II}} e^{-\mu l} \qquad (15).$$

*Equation (11) contains a common factor $\nu - \mu U$ which has been omitted. The equation is therefore also satisfied where $\nu - \mu U = 0$ for a fixed value of y. It can be shown that the original state of flow is neutral for a disturbance of this kind (Rayleigh, 1895).

From the continuity of the pressure, the following relations are obtained, substituting from equation (15)

$$bK_1^{II} - [2(\nu - \mu U_1) - b]\,K_2^{II} = 0,$$
$$[2(\nu - \mu U_2) + b]\,K_1^{II}e^{\mu l} + bK_2^{II}e^{-\mu l} = 0 \qquad (16).$$

If K_1^{II} and K_2^{II} do not vanish simultaneously—otherwise no wave motion would exist—the determinant of equation (16) must be equal to zero. This condition leads to an expression for the wave velocity c, since $c = \frac{\nu}{\mu} = \frac{L}{\tau}$,

$$(c - U_1)(c - U_2) + \frac{bL}{4\pi}(U_2 - U_1) - \frac{b^2L^2}{8\pi^2}\,\frac{1}{1 + \coth\frac{2\pi}{L}l} = 0.$$

According to equation (3a)

$$bl = U_2 - U_1.$$

Substituting this expression in the preceding equation and solving for c, it follows that

$$c = \frac{U_1 + U_2}{2} \pm \frac{U_2 - U_1}{2}\sqrt{1 - \frac{L}{l\pi} + \frac{L^2}{l^2\,2\pi^2\left(1 + \coth\frac{2\pi}{L}l\right)}} \qquad (17)$$

First, the case may be considered in which $\frac{2\pi}{L}l$ is so small that with sufficient accuracy $\coth\frac{2\pi}{L}l = \frac{L}{2\pi l}$. This approximation is correct within 1 per cent when $\frac{2\pi}{L}l < 0.2$ or $l < 0.03\,L$. For such a thin transitional layer,

$$c = \frac{U_1 + U_2}{2} \pm i\,\frac{U_2 - U_1}{2}\sqrt{\frac{1 - \frac{2\pi}{L}l}{1 + \frac{2\pi}{L}l}} \qquad (18).$$

These waves are always unstable, due to the wind shear, $U_2 - U_1$, through the transitional layer. The only difference from the case of a sharp discontinuity is the appearance of the factor,

$$\sqrt{\left(1 - \frac{2\pi}{L}l\right)\Big/\left(1 + \frac{2\pi}{L}l\right)},$$

in the imaginary term, because if $l = 0$,

$$c = \frac{U_1 + U_2}{2} \pm i\,\frac{U_2 - U_1}{2} \qquad (19).$$

In the general case of an arbitrary thickness of the transitional layer, instability occurs according to equation (17) when

$$l\frac{2\pi}{L}\left(l\frac{\pi}{L} - 1\right)\left(1 + \coth\frac{2\pi}{L}l\right) + 1 < 0.$$

This condition is satisfied when

$$\frac{2\pi}{L}l \leqq 1.2785,$$

or when

$$l = \leqq 0.204L \qquad (20).$$

Thus, waves whose length is less than about one-fifth of the thickness of the transitional layer are stable in spite of the wind shear.

In the absence of a stabilizing density discontinuity, waves at a sharp discontinuity of the velocity ("shearing waves") are always unstable, as can be seen from equation (19). If the sharp velocity discontinuity is replaced by a gradual transition, however, a limiting wave length exists according to equation (20), below which waves are stable in spite of the shear.

The surfaces of discontinuity that are observed in air or in water are all more or less gradual zones of transition. But the instability condition (20) will, as a rule, be satisfied except on a very diffuse frontal zone on which no cyclone formation can be expected. Since a wave motion of the dimensions of cyclone waves is very nearly horizontal, the width l of the transitional layer referred to in the preceding calculations would be represented approximately by the horizontal width of the frontal zone. It should be borne in mind in this connection that a frontal zone will be more diffuse in the surface layers than in the free atmosphere.

In the development of cyclone waves the stabilizing influences, which are the gravitational stability of stratification and the dynamic stability due to the earth's rotation, are counteracting the destabilizing effect of the velocity shear. Since the existence of a zone of transition decreases the destabilizing effect of the shear, the stabilizing effects must also decrease before unstable waves can develop. In the case of cyclone waves, the stabilizing effect of the stratification decreases with increasing wave length, because with increasing wave length the wave motion is tilted more toward a horizontal position. Therefore, the lower limit for the wave length of the unstable cyclone-waves should be larger the more diffuse the frontal zone.

Sufficiently long waves are stable again because with the increasing

dimensions of the wave disturbance, the effect of the dynamic stability due to the earth's rotation increases. Therefore, beyond a certain limiting wave length the shearing instability is overcompensated by the dynamic stability. Since the effect of a transitional layer is to reduce the shearing instability, it follows that the limiting wave length is smaller the wider the transitional zone. The influence of a transitional zone consists consequently in a narrowing of the region of unstable cyclone waves situated between the shorter waves, which are stable due to the stabilizing influence of the stratification, and the longer waves, which are dynamically stable.

This reduction of the region of wave lengths in which unstable waves are possible suggests that the formation of cyclones as frontal waves does not occur as frequently as might be expected from the theoretical results based on the assumption that the air-mass boundaries are mathematically sharp surfaces of discontinuity. The argument is admittedly not conclusive for a number of reasons, but it lends support to the opinion that other processes besides the wave mechanism may often produce cyclones.

Bjerknes and collaborators[1] have shown under certain simplifying assumptions that for sufficiently long waves the wind shear must be more than three times greater than the temperature difference at the frontal surface. In view of the relation (20), the factor three would presumably have to be replaced by a somewhat higher one.

Consider as an example a cyclone wave whose length is 2000 km at a frontal zone whose horizontal width is 100 km. With these figures the square root in equation (18) becomes 0.81, so that the magnitude of the unstable term is 20 per cent less than in the case of a sharp front, a deviation which is quite appreciable.

The type of motion associated with these unstable waves may be considered briefly. Writing the expression for the wave velocity, equations (17) or (18), in the abbreviated form

$$c = c_1 \pm ic_2 = \frac{U_2 + U_1}{2} \pm i\frac{\Delta U}{2}r,$$

where r stands for the square root in equation (17) or (18) and $\Delta U = U_2 - U_1$, the periodicity factor

$$e^{i(\mu x - \nu t)} = e^{i\mu(x - c_1 t)}\, e^{\pm \mu c_2 t}.$$

The constants K^{I}, K_1^{II}, K^{III} may be expressed by K_2^{II}, which will now be denoted by C. According to the first equation (16)

$$K_1^{II} = \left[\Delta U\,(1 \pm ir)\frac{\mu}{b} - 1\right]C = [\mu l\,(1 \pm ir) - 1]\,C.$$

From the first and second equations (15) it follows that

$$K^{\mathrm{I}} = \mu l\,(1 \pm ir)\,C,$$
$$K^{\mathrm{III}} = \left\{[(\mu l - 1) \pm ir\mu l]\,e^{2\mu l} + 1\right\} C.$$

In the following discussion only the positive sign will be chosen, as the negative sign refers to a wave whose amplitude decreases exponentially with time. Then, from equation (12)

$$v^{\mathrm{I}} = \mu l\,(1 + ir)\,Ce^{\mu c_2 t}\,e^{\mu y}\,e^{i\mu(x - c_1 t)}.$$

The constant C can be regarded as real without loss of generality. Since the real part of the expression for v^{I} must be a solution of the differential equations,

$$v^{\mathrm{I}} = C\mu l\,e^{\mu y}\,e^{\mu c_2 t}\,[\cos\mu(x - c_1 t) - r\sin\mu(x - c_1 t)],$$

or

$$v^{\mathrm{I}} = C\mu l\sqrt{1 + r^2}\,e^{\mu y}\,e^{\mu c_2 t}\cos[\mu(x - c_1 t) + \alpha^{\mathrm{I}}] \tag{21}$$

where $\tan\alpha^{\mathrm{I}} = r.$

Similarly

$$v^{\mathrm{II}} = C\sqrt{[(\mu l - 1)\,e^{\mu y} + e^{-\mu y}]^2 + r^2\mu^2 l^2 e^{2\mu y}}\,e^{\mu c_2 t}\cos[\mu(x - c_1 t) + \alpha^{\mathrm{II}}] \tag{22}$$

where

$$\tan\alpha^{\mathrm{II}} = \frac{r\mu l}{\mu l - 1 + e^{-2\mu y}},$$

and

$$v^{\mathrm{III}} = C\sqrt{[(\mu l - 1)\,e^{2\mu l} + 1]^2 + r^2\mu^2 l^2 e^{4\mu l}}\,e^{-\mu y}\,e^{\mu c_2 t}\cos[\mu(x - c_1 t) + \alpha^{\mathrm{III}}] \tag{23}$$

where

$$\tan\alpha^{\mathrm{III}} = \frac{r\mu l}{\mu l - 1 + e^{-2\mu l}}.$$

If the layer is sufficiently narrow compared to the wave length, $l \ll L$, $r \sim 1 - \mu l$, according to equation (18). Hence

$$\tan\alpha^{\mathrm{I}} = 1 - \mu l,$$
$$\tan\alpha^{\mathrm{II}} = \frac{1 - \mu l}{1 - \dfrac{2y}{l}},$$
$$\tan\alpha^{\mathrm{III}} = -(1 - \mu l).$$

These relations show that α^{I} is somewhat smaller than 45°. α^{III} is in the neighborhood of 135°, but somewhat larger. The phase difference of the transverse velocity between the two main layers is therefore slightly greater than 90°, or $\frac{L}{4}$, depending on the ratio of the wave length to the

width of the zone. In the case of a sharp surface of discontinuity, the phase difference between the transversal velocity components is exactly 90°. In the case of a gradual wind variation, the phase difference changes in a continuous fashion through layer II from its value in I to its value in III. Since the phase difference increases with increasing values of y, a phase retardation occurs in this direction throughout the transitional layer.

STABILITY AT A DISCONTINUITY OF THE WIND SHEAR

The solution of equations (6) and (7b) can also be applied to the discussion of the stability in the case of two fluid layers which are separated by a boundary at which the velocity is continuous while the shear $\frac{dU}{dy}$ changes abruptly (FIGURE 2). This type of wind distribution

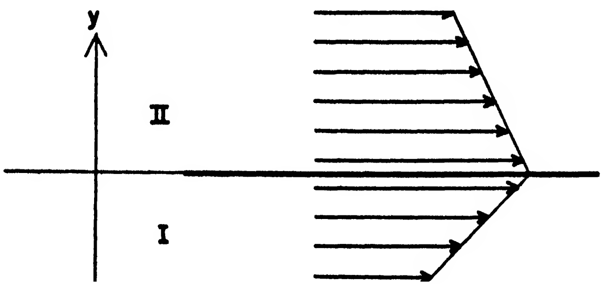


FIGURE 2. Discontinuity in wind shear between layers I and II.

is frequently found at the tropopause, at least in temperate latitudes. In order to show that waves at such a discontinuity of the wind shear are not unstable, both fluid layers will be assumed to have the same density so that gravitational stability does not exist. The y-axis may now be regarded as vertical. The effect of the earth's rotation can be neglected. Let the equation of the boundary be: $y = 0$; the wind velocity in the lower layer, $U^{I} = U_0 + b^{I}y$; in the upper layer, $U^{II} = U_0 + b^{II}y$; both layers extending to infinity. It follows, omitting the periodicity factor, $e^{i(\mu x - \nu t)}$, that

$$v^{\mathrm{I}} = K^{\mathrm{I}} e^{\mu y},$$
$$v^{\mathrm{II}} = K^{\mathrm{II}} e^{-\mu y}.$$

The perturbation pressures are, according to equation (10):

$$p^{\mathrm{I}} = i\frac{\rho}{\mu}[b^{\mathrm{I}} + (c - U^{\mathrm{I}})\mu] K^{\mathrm{I}} e^{\mu y},$$

$$p^{\mathrm{II}} = i\frac{\rho}{\mu}[b^{\mathrm{II}} - (c - U^{\mathrm{II}})\mu] K^{\mathrm{II}} e^{-\mu y}.$$

From the boundary conditions that the vertical velocity and the pressure are continuous at $y = 0$, it follows that

$$c = U_0 + \frac{L}{2\pi}\frac{b^{\mathrm{II}} - b^{\mathrm{I}}}{2} \qquad (24).$$

The velocity of the wave is equal to the arithmetic mean of the wind velocities in the first and in the second layer at a distance $\frac{L}{2\pi}$ from the boundary surface. The wave velocity is entirely convective, without a dynamic term, and no shearing instability occurs. A more general expression which includes the effects of a wind discontinuity and a density discontinuity has been given elsewhere (Haurwitz[2]).

The result is of meteorological interest insofar as the assumed wind distribution resembles that frequently found at the tropopause. Throughout the troposphere the westerlies of temperate latitudes increase, due to the temperature decrease polewards; in the stratosphere they decrease again, without a sharp discontinuity of the wind at the tropopause. Owing to the absence of shearing instability at such a discontinuity of the first order, waves at the tropopause will not form spontaneously as they may do at frontal surfaces under favorable conditions. Waves at the tropopause will be, in general, rather of a secondary nature, produced by waves at the polar front.

DISCUSSION OF THE PAPER

Prof. J. Holmboe (*University of California at Los Angeles, Calif.*):

Prof. Haurwitz is to be complimented on his mathematical simplification of a very difficult problem. There exists a lack of agreement between several writers regarding the destabilising effect of a shearing motion. Rayleigh's classical result is that a homogeneous straight current is unstable when the velocity profile has inflections (the velocity gradient changes sign), and stable in all other cases. Hoiland[3] has recently shown by a different method, using the circulation theorem, that stable waves are impossible in a straight current with arbitrary two-dimensional shear. Hoiland's result was derived from the complete equation of motion without linearization and is also supported by the experimental fact that laminar flow in a pipe breaks down beyond the critical Reynolds number, although the velocity profile is parabolic in this case. Therefore, I believe that Hoiland's result was correct and that Rayleigh's criterion perhaps might be due to the simplifications introduced by using linear equations and dropping the higher-order terms.

Reply by Prof. B. Haurwitz *(communicated after the conference)*:

In the discussion of the paper reference was made by Prof. Holmboe to a paper by Hoiland.[3] In this paper it is deduced from qualitative considerations based on the circulation theorem, that shearing instability must exist for any wave length, even if the wind discontinuity is replaced by a zone of gradual transition, whereas Lord Rayleigh's results, on which the present paper is based, are said to refer to a different type of waves.

Hoiland's qualitative arguments in favor of unstable waves even in the case of a gradual wind change have a considerable degree of plausibility. In order to make them convincing, however, it would be necessary to show by mathematical methods that unstable types of waves are possible even under the condition when the wave types discussed here are stable Until such a mathematical deduction of the wave type postulated by Hoiland is forthcoming, there does not seem to be any reason to doubt the applicability of the results presented here to the cyclone theory.

Dr. C. L. Pekeris *(Columbia University, New York, N. Y.)*:

A "physical" explanation of Rayleigh's theorem was given by G. I. Taylor in 1915 (Phil. Trans. Roy. Soc., A **215**: 23). Taylor shows that, under the assumption of conservation of vorticity in eddy motion, the gain of x-momentum at each level during the breakdown of the laminar flow is proportional to d^2U/dz^2. If d^2U/dz^2 is of the same sign throughout an inviscid fluid, the whole body of the fluid will be gaining momentum. Since the walls cannot communicate the required momentum, the breakdown is therefore impossible. If the viscosity is finite, there is no difficulty because the walls can supply the required momentum. Taylor's considerations offer a physical explanation of Rayleigh's theorem and also indicate that it is not limited in its application to infinitesimal disturbances. But in the case of parabolic viscous flow, it has been established both theoretically and experimentally that the motion is stable for infinitesimal disturbances.

The main conclusion of Prof. Haurwitz's paper is an instance of a well-known dynamical theorem that a feature of wave motion which is due to the existence of a discontinuity in the properties of the medium will persist even if the change is continuous, provided this transition takes place in a space that is small in *comparison with the wave length*. When the transition takes place in a space interval that is long in comparison with the wave length, the feature disappears. The reason for association between the space over which the transition takes place and the wave length can be looked for in the fact that in the case of a sharp transition the distance normal to the plane of discontinuity over which the free wave motion is appreciable is of the order of a wave length.

REFERENCES

1. **Bjerknes, V., & coworkers**
   1932. Physikalische Hydrodynamik. Berlin. P. 613.
2. **Haurwitz, B.**
   1931. Veröff. Geophys. Inst. Leipzig, 2d ser. **5**: 68.
3. **Hoiland, E.**
   1939. Arch. f. Math. og Naturvidenskab., B. **42**: 68.
4. **Rayleigh, Lord**
   1895. Proc. London Math. Soc. **27**: 5.
5. 1896. Theory of sound, 2d ed., Vol. II, Chap. XXI.
6. **Taylor, G. I.**
   1931. Effect of variation in density on the stability of superposed streams of fluid. Proc. Roy. Soc. London **132**: 499–523.

ON THE RATIO BETWEEN HEAT CONDUCTION FROM THE SEA SURFACE AND HEAT USED FOR EVAPORATION*

By H. U. Sverdrup

Scripps Institution of Oceanography, University of California, La Jolla, California

Since the evaporation from the oceans by means of the energy equation was first calculated[11] in 1915, numerous attempts have been made to use this procedure and to apply it to other problems of similar nature. If no energy flows through the solid boundaries of a body of water the energy equation takes the form $Q_a = Q_r - Q_\theta$, where Q_a represents the energy which is given off from the water surface to the air, Q_r is the net radiation received, that is, the difference between the incoming radiation from sun and sky and the outgoing long-wave radiation of the surface, and Q_θ is the heat used in changing the temperature of the water.

In the above form the energy equation simply states that the difference between energy received and used for changing the temperature must be given off to the atmosphere. In order to find the evaporation it is necessary to consider that the energy given off to the air is composed of two terms, one representing the energy used for evaporation, Q_e, and one representing that which is conducted to the air as sensible heat, Q_h. Thus, $Q_a = Q_e + Q_h$. Introducing the ratio, $R = \frac{Q_h}{Q_e}$, and writing $Q_e = LE$, where L is the latent heat of evaporation and E is the evaporation, one obtains

$$E = \frac{Q_r - Q_\theta}{L\,(1 + R)} \qquad (1),$$

where the units of E depend upon the units in which the other terms are measured.

Since Q_r and Q_θ can be obtained accurately from adequate observation, it is important also to determine R. In his computations, Schmidt assumed values of R ranging from 2.3 to 0.3. Ångström[1] pointed out that these values were probably too high and, from a consideration of the processes of heat conduction and evaporation, he concluded that on an average R was probably about 0.1. This value was used by Mosby[8] when he revised Schmidt's figures.

Meanwhile Bowen[2] had established an expression for R which is

*Contributions from the Scripps Institution of Oceanography, New Series, No. 202.

strictly correct only under the conditions considered by him. This expression, "the Bowen ratio," has been applied to atmospheric conditions, particularly by Cummings and Richardson,[6] Richardson,[9] and Cummings[4,5] who have assumed that the Bowen ratio would be exactly valid in this case as well. An analysis of the general validity of the Bowen ratio will be attempted here.

The two fundamental equations from which Bowen starts can be written

$$Q_h = D_2 c_p \rho(\vartheta_1 - \vartheta_2)/l \qquad (2),$$

$$Q_e = D_1 L (\rho_1' - \rho_2')/l \qquad (3),$$

where D_2 is the "diffusion coefficient" for heat energy and D_1 is the diffusion coefficient of water vapor through air. These coefficients have been found to differ only by a few per cent, as predicted by the kinetic gas theory. Furthermore, c_p is the specific heat at constant pressure, ρ the density of the air, ϑ_1 and ϑ_2 the air temperature at the two faces of a space of length l, ρ_1' and ρ_2' the densities corresponding to the vapor pressures, and L is the latent heat of vaporization.

Bowen discusses three different cases, but it is sufficient here to consider case II, in which it is assumed that heat and water vapor are carried away by processes of diffusion only. On these assumptions, the ratio, Q_h/Q_e is

$$R_B = \frac{D_2}{D_1}\frac{c_p\rho}{L}\frac{\vartheta_1 - \vartheta_2}{\rho_1' - \rho_2'} \qquad (4).$$

Introducing the sufficiently accurate expression

$$\rho' = 0.621\,\rho\frac{e}{p} \qquad (5),$$

where e is the vapor pressure and p is the atmospheric pressure; putting $\vartheta_1 = \vartheta_w$, the temperature of the water surface, $\vartheta_2 = \vartheta_a$, the temperature of the air at any height above the water, $e_1 = e_w$, the vapor pressure at the water surface, and $e_2 = e_a$, the vapor pressure in the air at the height where the temperature was measured; and using the numerical values $D_2 = 0.181\ f(\vartheta)$, $D_1 = 0.206\ f(\vartheta)$, $c_p = 0.241$, and $L = 585$, Bowen obtains

$$R_B = 0.58\frac{\vartheta_w - \vartheta_a}{e_w - e_a}\frac{p}{1000} \qquad (6).$$

It is here assumed that Q_h and Q_e are independent of the height above the water surface; that is, that stationary conditions exist. Furthermore, the assumption is made that the transfer of heat energy and water vapor takes place by ordinary diffusion, in which case the corresponding

coefficients are functions of temperature only. These functions are similar and the ratio D_2/D_1 is therefore independent of temperature and also independent of the distance from the water surface.

The fundamental equations from which Bowen starts apply to transfer of heat and water vapor by processes of ordinary diffusion. It might therefore appear that the computed ratio R_B is valid only when the air is at rest or in laminar motion, but, as stated by Bowen, one can expect "that heat losses by evaporation and diffusion and by conduction will follow the same laws and will be affected in the same way by convection" (turbulence). This implies that the ratio between the heat losses will be independent of the state of turbulence, or, that the ratio, R_B, will be valid regardless of the character of the air motion. Recent results suggest, however, that when examining the transfer of heat energy it may be necessary to take into account the effect of processes of radiation upon the transfer of heat and the effect of spray upon evaporation. If such processes have to be considered it can be expected that heat losses by conduction, and by evaporation and diffusion, will be affected differently by turbulence and that the Bowen ratio may have to be modified.

Let us first examine the effect of heat transfer by radiation when the air is at rest or in laminar motion. In a moist atmosphere a radiant flux of heat is directed from regions of higher to regions of lower temperature (Brunt, 1934, p. 119) This flux is dependent upon the selective emission and absorption of long-wave radiation by the water vapor, and is not present in dry air because dry air does not emit or absorb long-wave radiation. The effect of this flux of radiation can be described by means of a coefficient which Brunt calls radiative diffusivity The radiative diffusivity, K_R, has the same dimensions (cm²/sec) as the diffusion coefficient for heat energy, D_2, and when problems of diffusion of heat energy in moist air are considered, must be added to D_2. Doing this, equation (6) is altered to

$$R = \frac{D_2 + K_R}{D_1} 0.66 \frac{\vartheta_w - \vartheta_a}{e_w - e_a} \frac{p}{1000} \tag{7}$$

Thus, the ratio R is greater than R_B and the increase depends upon the relative magnitudes of D_2 and K_R. Unfortunately, very little is known about the value of K_R near a boundary surface. At distances above the ground greater than about 40 meters, K_R has a value, according to Brunt, of the order of 10^3cm²/sec and is thus about 10^4 times greater than D_2, but it seems certain that K_R must decrease rapidly when approaching a boundary surface. The very great value computed by

Brunt at a moderate height strongly suggests that near a boundary surface K_B may be of the same order of magnitude as D_2 (about 0.2 cm^2/sec) and that in the absence of turbulence R is many times greater than indicated by Bowen's formula. Theoretical and experimental examination of this question seems desirable.

In the presence of turbulence, neglecting radiative diffusivity, the equations for transfer by eddy diffusivity of heat energy and of energy in the form of latent heat take the form

$$Q_h = -\frac{\mu_e}{\rho} c_p\rho\,((d\vartheta/dz) + \gamma) \qquad (8),$$

$$Q_e = -\frac{\mu_e}{\rho} L\, d\rho'/dz \qquad (9),$$

where μ_e/ρ is the eddy diffusivity and γ is the adiabatic lapse rate, 1.0C.° per 100 meters.

In a steady state Q_h and Q_e are independent of the height above the water surface and represent the quantities indicated by the same symbols in equations (2) and (3). With the above numerical value, the equation

$$R = 0.66\,\frac{p}{1000}\left(\frac{(d\vartheta/dz) + \gamma}{de/dz}\right) \qquad (10),$$

can therefore be considered valid within the turbulent layer. If this equation were valid from the very water surface, the differentials could be replaced by differences and

$$R = 0.66\,\frac{p}{1000}\,\frac{\vartheta_w - \vartheta_a - \gamma h}{e_w - e_a} \qquad (11),$$

where h is the height at which the air temperature is measured, would be an exact equation for the ratio Q_h/Q_e.

This expression for R differs little from equation (6) if $(\vartheta_w - \vartheta_a)$ is much greater than γh. This is usually the case and the Bowen ratio, R_B, is therefore applicable, as pointed out by Bowen, in the presence of turbulence if other processes can be disregarded.

In the turbulent layer near the ground the effect of radiative diffusivity can probably be disregarded because at moderate and high wind velocities the eddy diffusivity is many times greater than the estimated radiative diffusivity, but observations of temperature and vapor pressure directly above a water or snow surface indicate the existence of a boundary layer through which the transfer of energy takes place by processes of molecular diffusion and in which radiative diffusivity may be important. This boundary layer is described by Brunt[3] (p. 262) as a time-mean phenomenon:

"Water-vapor produced by evaporation at a liquid surface spreads outward through the laminar layer by molecular diffusion, while the air within that layer is from time to time exchanged by the action of eddies which penetrate it."

Above the boundary layer the transfer of heat and water vapor takes place by eddy diffusion as expressed by equations (8) and (9). As a first approximation it may be assumed that, near a surface which is hydrodynamically rough, the eddy diffusivity is a linear function of the height above the surface (Rossby[10]), in which

$$\mu_e/\rho = k_0 w^*(z + z_0) \tag{12}$$

where $k_0 = 0.4$ is von Kármán's constant and w^* is the friction velocity ($w^* = \sqrt{\tau_0/\rho}$, where τ_0 is the stress of the wind). Assuming this value of the eddy diffusion above the boundary layer and introducing the ordinary diffusion coefficient in the boundary layer,[13] the upward transfer of water vapor can be represented as a function of the difference $(e_w - e_a)$ and, following a similar procedure, the heat transfer can similarly be expressed as a function of $(\vartheta_w - \vartheta_a)$. Assuming that radiative diffusivity can *not* be neglected in the boundary layer, one obtains

$$R = 0.66 \frac{p}{1000} \frac{\vartheta_w - \vartheta_s - \gamma h}{e_w - e_s} \frac{\ln \dfrac{z + z_0}{d + z_0} + \dfrac{k_0 w^* d}{D_1}}{\ln \dfrac{z + z_0}{d + z_0} + \dfrac{k_0 w^* d}{D_2 + K_R}} \tag{13}$$

where d is the thickness of the boundary layer. The above treatment assumes a discontinuity in the value of the diffusivity at the top of the boundary layer and in the first approximation this assumption appears satisfactory.

With $z = 600$ cm, $z_0 = 0.6$ cm, $d = 0.25$ cm, $w^* = 16$ cm/sec (corresponding to a wind velocity of about 6 m/sec at a height of 6 m), $k_0 = 0.4$, $D_1 = 0.206$ cm^2/sec, $D_2 = 0.181$ cm^2/sec, and $K_R = D_2$, one obtains a ratio, Q_h/Q_e, which is 1.30 times greater than computed by means of equation (11). Thus, Bowen's formula may give too small values of R when applied to atmospheric conditions, but this conclusion is based on inadequate knowledge as to the radiative diffusivity directly above the water surface. A few observations suggest, however, that radiative diffusivity must be considered.[12]

The application of Bowen's formula to atmospheric conditions will be further invalidated if the evaporation from a water surface is greatly increased at wind velocities high enough to carry spray to the air. Montgomery[7] has suggested that this process is of great importance and

that as a result of evaporation from spray the evaporation will not remain proportional to the wind velocity, as is nearly the case when transfer by eddy conductivity only is considered, but will increase more rapidly at higher wind velocities. In a formal manner one can state that due to the effect of spray the evaporation will not be proportional to the wind velocity, w, but to w^n, where n is greater than 1.0. Heat conduction, on the other hand, will not be altered by the presence of spray and will remain proportional to the wind velocity. Therefore, the Bowen ratio must be multiplied by a factor, w^{1-n}; that is, at higher wind velocities the Bowen ratio, R_B, gives too high values of R.

The above analysis leads to the conclusion that the Bowen ratio, R_B, when applied to atmospheric conditions, gives only an approximately correct value of the ratio, $R = Q_h/Q_e$. The closeness of the approximation depends upon the importance of radiative diffusivity, the character of the turbulence near the ground, and the effect of spray upon evaporation. At present, none of the processes involved is well enough understood to make possible exact determinations of the ratio R under varying conditions. In the computation of evaporation, the term RL (equation 1) has the character of a correction term, and uncertainty as to the exact value of R does not, therefore, seriously impair the usefulness of the energy equation for determination of evaporation.

DISCUSSION OF THE PAPER

Prof. A. F. Spilhaus (*New York University, New York, N. Y.*):

Meteorologists do not use the Bowen ratio at all, but it is up to them to convince hydrologists that it is unsound and to point out better methods.

Prof. B. Haurwitz (*Massachusetts Institute of Technology, Cambridge, Mass.*):

The value of K_R for the radiative transfer of heat is certainly much larger than 10^3 c.g.s. units according to the new values for the absorption coefficients of water vapor. The above value was computed by Brunt under the assumption that a layer containing 0.3 mm. of precipitable water absorbs completely the so-called W-radiation. Such a layer would have a thickness of 40 meters next to the ground. But according to Elsasser, with the new absorption coefficients the amount of precipitable water should be very considerably larger, which would increase K_R at least by a factor of 10. Moreover, the theory is based on the assumption that the layers of complete absorption of W-radiation are thin. This assumption may be considered correct as long as a layer of 40 meters (0.3 mm. of precipitable water) absorbs completely. But if the layer is ten times thicker, as it should be, at least, in view of the newer absorption coefficients, the whole analogy between radiative and turbulent transfer of heat breaks down. The thickness of the layers considered necessary by Elsasser are too thick for use of K_R in a shallow layer of the ocean.

Dr. H. Wexler (*U. S. Weather Bureau, Washington, D. C.*):

I agree with Professor Haurwitz that the method of using K_R for layers of ocean is inapplicable. It is suggested instead that Elsasser's Radiation Chart (2d ed.) be used to find the radiative heating. Use of Brunts' formula is extremely uncertain in view of the thin layer involved and highly selective character of radiation.

Mr. W. C. Jacobs (*U. S. Weather Bureau, Washington, D. C.*):

The use of K_R affects my own computations seriously only in the expressions for Q_s and Q_e, but not the evaporation.

Prof. N. W. Cummings (*San Bernadino Valley Junior College, Calif.*) communicated a letter to the conference, stating that although he and Dr. Sverdrup were in agreement on most points concerning the Bowen ratio, he would like to outline some points of difference in their views, as follows (read by Professor Spilhaus):

Dr. Sverdrup lays great stress on the transfer of heat by radiation, and urges further investigation of the question. In this, he is, of course, correct. Nevertheless, we must avoid any exaggerated idea of the probable magnitude of the source of error he has in mind. Moreover, we must recognize the intimate relation between radiative diffusivity and the long-wave radiation which is subtracted from the sun and sky radiation in order to get Dr. Sverdrup's Q_r. In the investigations made by the present writer, radiation has been dealt with in a manner which, in the light of Dr. Sverdrup's discussion, may be described as follows:

His quantity K_R has not been entirely neglected, but has been treated as if it were a part of the outgoing long-wave radiation just mentioned. This treatment may not be strictly correct, but it is certainly not as bad as a total neglect of this energy component. The results of the experiments of Cummings and Richardson strongly indicate that Dr. Sverdrup's suggestion of a manyfold error is excessive. In fact, the error seems imperceptible. It should be noted that the experiments were undertaken for the specific purpose of providing an over-all test of the general procedure which involves an application of the Bowen equation.

It is difficult to understand the remarks about spray, contained in the latter part of the paper. There is no obvious reason for supposing that spray constitutes an exception to the generalization pointed out by Bowen, that the transport of mass and of sensible heat follow the same laws. In fact, it seems highly probable that the case in which spray occurs can be analyzed by a method similar to Bowen's case (III).

REFERENCES

1. **Ångström, Anders**

1920. Applications of heat radiation measurements to the problems of the evaporation from lakes and the heat of convection at their surfaces. Geog. Ann. Stockholm, Heft 3. 16 p.

2. **Bowen, I. S.**

1926. The ratio of heat losses by conduction and by evaporation from any water surface. Phys. Rev. **27**: 779 787.

3. **Brunt, David**

1934. Physical and dynamical meteorology. New York and London. 411 p.

4. **Cummings, N. W.**

1936. Evaporation from water surfaces: status of present knowledge and need for further investigations. Trans. Am. Geophys. Un. **1936**: 507–509.

5. 1940. The evaporation-energy equations and their practical applications. Trans. Am. Geophys. Un. **1940**: 512 522.

6. **Cummings, N. W., & Richardson, Burt**

1927. Evaporation from lakes. Phys Rev. **30**: 527 534.

7. **Montgomery, R. B.**

1940. Observations of vertical humidity distribution above the ocean surface and their relation to evaporation. Papers in Phys. Oceanogr. and Meteor. **7** (4). 30 p.

8. **Mosby, Hakon**

1936. Verdunstung und Strahlung auf dem Meere. Ann. d. Hydrogr. u. Mar. Meteor. **64**: 281–286.

9. **Richardson, Burt**

1931. Evaporation as a function of insolation. Trans. Am. Soc. Civ. Engin. **95**: 996–1011, discussion 1012–1019.

10. **Rossby, C.-G.**

1936. On the frictional force between air and water and on the occurrence of a laminar boundary layer next to the surface of the sea. Papers in Phys. Oceanogr. and Meteor. **4** (3). 20 p.

11. **Schmidt, Wilhelm**

1915. Strahlung und Verdunstung an freien Wasserflachen; ein Beitrag zum Wärmehaushalt des Weltmeers und zum Wasserhaushalt der Erde. Ann. d. Hydrogr. u. Mar. Meteor. **43**: 111–124.

12. **Sverdrup, H. U.**

1936. The eddy conductivity of the air over a smooth snow field. Geofysiske Publikasjoner. **11** (7). 69 p.

13. 1937. On the evaporation from the oceans. Jour. Marine Research **1**: 3–14.

GENERALIZATION FOR CYLINDERS OF PRANDTL'S LINEAR ASSUMPTION FOR MIXING LENGTH*

By R. B. Montgomery

New York University and Woods Hole Oceanographic Institution

INTRODUCTION

In atmospheric and oceanic boundary-layer problems certain crucial quantities have defied direct measurement. One of these is the shearing stress between the air and the ground or between the air and the ocean surface. This has been computed indirectly: on the one hand from the frictional retardation of the wind as manifested in the transport of air in the direction of the horizontal pressure gradient (across the isobars); and on the other hand from the slope of the sea surface produced by the wind in constricted parts of the ocean and from the movement of sea ice. Another such quantity is evaporation from natural land and water surfaces, which has been estimated from the thermal energy used in vaporization as computed from the other terms in the heat balance for the ground or water. The values resulting from these various methods are not of high accuracy.

A method sometimes used in the calculation of these quantities utilizes the controlled experiments on turbulent flow which have been obtained in hydraulic and aerodynamic laboratories. The experiments which lend themselves most readily to this application are those on the flow of fluids through long circular cylinders. Due to the relative simplicity of the model, the velocity distribution and axial pressure gradient can be measured with great accuracy, and the latter gives the shearing stress at any distance from the axis. Except near the entrance the mean velocity at a point depends on one coordinate alone, the distance from the axis. In this important respect the flow is similar to a horizontally uniform wind, in which the mean velocity depends only on the distance from the ground, a boundary essentially in the form of a plane of infinite extent.

The infinite plane and the infinitely long cylinder of small radius are nevertheless boundaries of definitely different geometric pattern. In applying to one the results from the other it is therefore necessary to be

*Contribution No. 252 from Woods Hole Oceanographic Institution

sure that the theory involved can be applied satisfactorily to both models.

The approach followed here is the very fruitful one due primarily to Prandtl and von Kármán. By means of their general theory the quantitative observations of the complicated pattern of turbulent flow can be expressed largely in terms of a single constant, the universal turbulence constant. The atmospheric applications give the shearing stress and evaporation as proportional to the square of this constant. The value of the constant obtained experimentally depends, however, on the particular form of the general theory employed. Two previous forms of the general theory, one giving the constant as 0.40 and the other as 0.31, are examined below and found unsatisfactory. A third form, which is apparently in accord with the observations, is advanced here and yields 0.45 as the value of the universal turbulence constant.

GENERAL CONSIDERATION OF TURBULENT FLOW IN A CIRCULAR CYLINDER

For the steady mean flow of an incompressible homogeneous fluid in a long circular cylinder the mean velocity is everywhere parallel to the axis and is therefore expressed completely by the one component u. The mean velocity varies only with distance z from the wall, or distance r from the axis. In terms of the radius r_0 of the cylinder,

$$z + r = r_0.$$

The *relative distance from axis* may be written as

$$k = \frac{r}{r_0} = 1 - \frac{z}{r_0} \tag{1}.$$

The effective shearing stress τ across any concentric cylindrical surface varies linearly with distance from the axis, which follows from the balance between normal and shearing stresses on any concentric fluid cylinder. If τ_0 is the wall stress,

$$\tau = \tau_0 k \tag{2}.$$

The *friction velocity*, defined as

$$u_* = \sqrt{|\tau|/\rho}, \qquad u_{*0} = \sqrt{|\tau_0|/\rho} \tag{3},$$

where ρ is density, therefore varies as

$$u_* = u_{*0}\sqrt{k} \tag{4}.$$

The effective shearing stress is composed of two terms,

$$\tau = \mu \frac{du}{dz} - \rho\, \overline{w'u'} \tag{5},$$

where μ is dynamic viscosity and $\overline{w'u'}$ is the mean product of the instantaneous velocity deviations in the directions normal to the boundary and parallel to the axis. Except in the vicinity of the boundary the first term is negligible for well-developed turbulent flow.

In order to evaluate the second term, the Reynolds stress, Prandtl[7] has introduced the *mixing length* l, which may be defined by

$$-\rho\,\overline{w'u'} = \rho\, l^2 \frac{du}{dz}\frac{du}{dz} \qquad (6).$$

Except near the boundary, then,

$$u_* = l\frac{du}{dz} = -\frac{l}{r_0}\frac{du}{dk} \qquad (7).$$

Combining equations (4) and (7) gives

$$\frac{l}{r_0} = -u_{*0}\frac{\sqrt{k}}{du/dk} \qquad (8),$$

by means of which the mixing length can be computed from the observed velocity distribution and wall stress. For sufficiently high Reynolds numbers Nikuradse[5,6] has found by this means that the *relative mixing length* l/r_0 depends only on the relative distance from axis. His empirical equation,

$$l/r_0 = 0.14 - 0.08\,k^2 - 0.06\,k^4 \qquad (9),$$

identical for both smooth and rough pipes, is shown graphically in FIGURE 1.

Although equation (8) is not always applicable close to the wall, it may be integrated to give the distribution of *velocity defect* $u_m - u$, where u_m is the mean velocity at the axis, throughout the rest of the cylinder,

$$\frac{u_m - u}{u_{*0}} = \int_0^k \frac{\sqrt{k}}{l/r_0}\,dk \qquad (10).$$

It has been found experimentally that for well-developed turbulent flow this ratio $(u_m - u)/u_{*0}$ is apparently a universal function of the relative distance from axis, the function extending close to the wall (von Kármán, 1934, p. 6). Some experimental points showing this function are plotted in FIGURE 2.

By studying the distribution of velocity defect instead of the distribution of velocity itself, one avoids the complicated conditions close to the boundary, where also the observations are inherently less accurate. Furthermore, the universal character found empirically for the velocity defect indicates that the central region of a cylinder offers the simplest proving ground for theory.

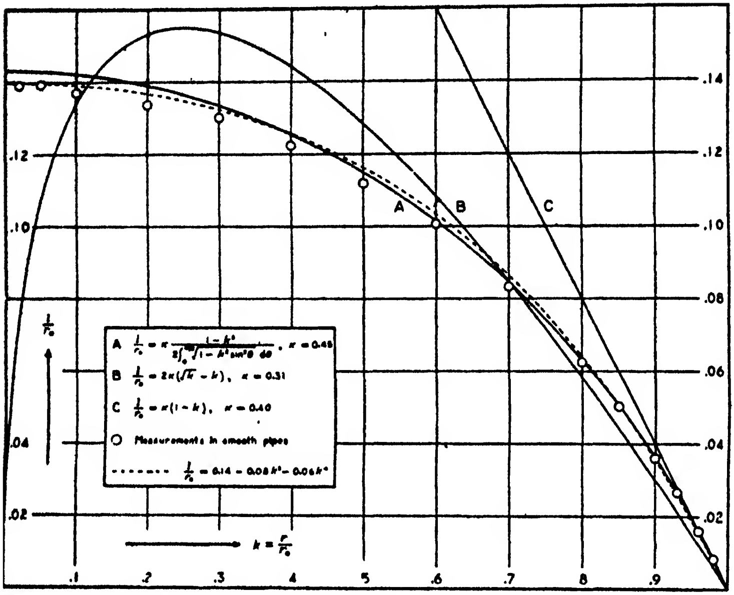


FIGURE 1. Relative mixing length as a function of the relative distance from axis of a circular cylinder.

A theory for the velocity distribution may be expressed in terms of the mixing length, and each of the three variations below introduces the universal turbulence constant. The satisfactory theory for the distribution of mixing length must not only be applicable for an infinite plane boundary but also agree with equation (9) for a circular cylinder. Or better, since the empirical equation (9) was determined from small differences of observed velocities, when inserted in equation (10) the mixing-length theory should give a function in close agreement with the directly observed distribution of the ratio $(u_m - u)/u_{*0}$.

LINEAR DISTRIBUTION OF MIXING LENGTH

Prandtl (1932, p. 189) found that in the immediate vicinity of a solid boundary the mixing length is, very nearly, simply a linear function of the shortest distance from the boundary. The proportionality factor is the *universal turbulence constant* κ. With an additive term in the form introduced by Rossby and Montgomery (1935, p. 5)

$$l = \kappa\,(z + z_0) \qquad (11).$$

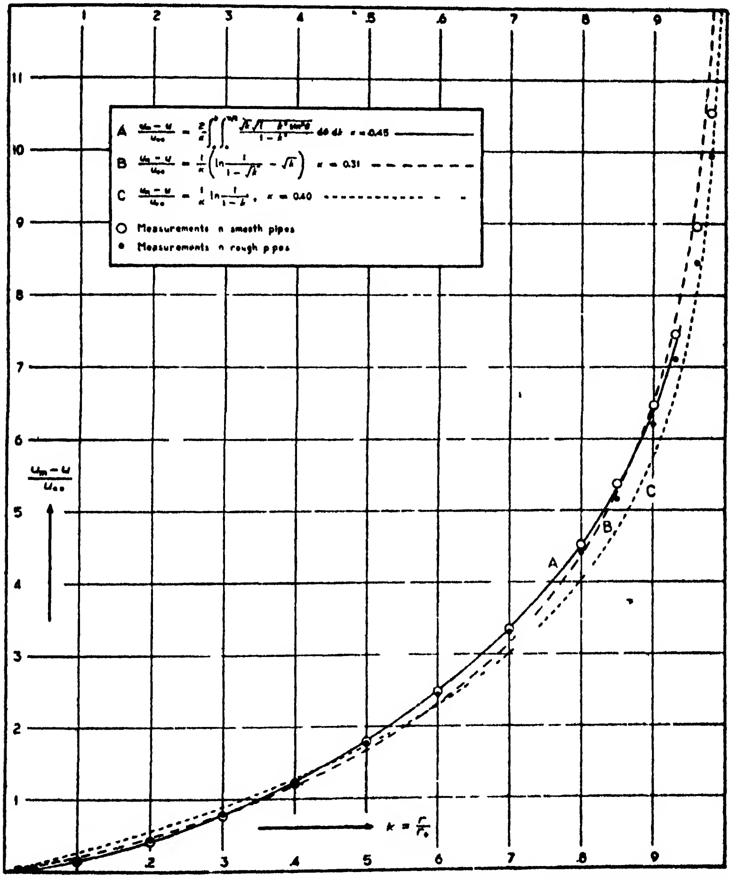


FIGURE 2. Ratio of velocity defect to friction velocity as a function of the relative distance from axis of a circular cylinder.

The length z_0, which for a hydrodynamically rough boundary has been found to depend on the nature of the boundary alone, may be called the *roughness length*.

From equation (7) it follows that

$$\frac{du}{u_*} = \frac{1}{\kappa}\,\frac{dz}{z + z_0}.$$

For a boundary layer within which the variation of friction velocity is

negligible this gives on integration, assuming the mean velocity to vanish at $z = 0$,

$$\frac{u}{u_{*0}} = \frac{1}{\kappa} \ln \frac{z + z_0}{z_0} \qquad (12).$$

Except for the thin layer next to the boundary where z_0 is comparable with z, the last equation may be written

$$\frac{u}{u_{*0}} = \frac{1}{\kappa} \ln \frac{z}{z_0} \qquad (13),$$

given first by Prandtl.[8] This logarithmic distribution of wind above the ground agrees with observation within experimental error when the air is in neutral hydrostatic equilibrium (Prandtl, 1932; Lettau, 1939, p. 72).

The assumption of linear distribution of mixing length and the disregard of the variation of friction velocity are in definite disagreement with conditions in a circular cylinder except within a boundary layer extending perhaps a tenth of the distance to the axis. If one nevertheless applies the last equation all the way to the axis, it gives

$$\frac{u_m - u}{u_{*0}} = \frac{1}{\kappa} \ln \frac{1}{1 - k} \qquad (14).$$

Surprisingly enough, this agrees rather well with the observed distribution. The reason that the two false assumptions lead to such a good result is that the basic differential equation (7) contains the ratio of friction velocity to mixing length, both of which, by virtue of the assumptions, are increasingly too large on approaching the axis. The two errors therefore cancel to some extent. It may be noted that the resulting expression gives finite shear at the axis, which is physically impossible.

A value of 0.40 for the universal turbulence constant gives the best agreement between this theoretical expression and observation (Rouse, 1938, p. 244). Curves C in FIGURES 1 and 2 show how good the agreement is.

The generally accepted value for the universal turbulence constant is 0.40, derived from a somewhat different treatment of the problem—using actual velocity instead of velocity defect—but still involving constant friction velocity and linear increase of mixing length well into the central portion of the cylinder (von Kármán, 1934, p. 9). For application over a plane boundary this determination is seen to be unsatisfactory.

MIXING LENGTH IN TERMS OF FIRST AND SECOND DERIVATIVES OF MEAN VELOCITY

By stipulating similarity of the pattern of turbulent flow, von Kármán (1930, p. 63) concluded that the mixing length is given by

$$l = \kappa \left| \frac{du}{dz} \Big/ \frac{d^2u}{dz^2} \right| \tag{15},$$

where κ is again the universal turbulence constant. From equation (7) it follows that

$$u_* \frac{d^2u}{dz^2} \Big/ \left(\frac{du}{dz}\right)^2 = \kappa.$$

For a boundary layer within which the variation of friction velocity is negligible this gives on integration

$$u_{*0} \Big/ \frac{du}{dz} = \kappa\,(z + \text{const.}).$$

The left side is the mixing length, so equation (15) reduces to equation (11) and therefore this assumption for mixing length is applicable for an infinite plane boundary.

At the axis of a small circular cylinder the shear vanishes but the second derivative of velocity is finite. According to equation (15) the mixing length would therefore vanish, which is contrary to observation. It is thus seen at once that this assumption is not suitable for flow in circular cylinders.

Taking into account the variation of friction velocity and introducing the relative distance from axis, the differential equation for the velocity distribution becomes

$$-u_{*0} \frac{d^2u}{dk^2} \Big/ \left(\frac{du}{dk}\right)^2 = \frac{\kappa}{\sqrt{k}} \tag{16}.$$

At the wall the shear is relatively great, hence as a boundary condition for the first integration the shear is assumed infinite at the wall, giving

$$-\frac{1}{u_{*0}} \frac{du}{dk} = \frac{1}{2\kappa} \frac{1}{1 - \sqrt{k}} \tag{17}.$$

This gives finite shear at the axis, so it is valid neither at the axis nor at the wall. According to this relation the relative mixing length is

$$l/r_0 = 2\kappa\,(\sqrt{k} - k) \tag{18}.$$

It vanishes at the axis and at the wall, and is maximum at a quarter of the distance from axis to wall.

Integration of equation (17) gives, as the equation for the velocity defect,

$$\frac{u_m - u}{u_{*0}} = \frac{1}{\kappa}\left(\ln \frac{1}{1 - \sqrt{k}} - \sqrt{k}\right) \tag{19}$$

In order to fit this to the observed distribution of velocity defect von Kármán (1934, Fig. 8) used 0.31 for the universal turbulence constant. With this value, equations (18) and (19) are drawn in FIGURES 1 and 2 respectively as curves B.

A NEW HYPOTHESIS FOR THE MIXING LENGTH

It is desirable to seek a more generally applicable assumption for the mixing length. For a plane boundary this assumption should reduce, as does von Kármán's, to a linear distribution.

Prandtl's geometrical assumption for mixing length can be generalized so that the dependence is not merely on the shortest distance to the boundary. It is reasonable to expect that at a given shortest distance to the boundary the mixing length is, for instance, less in a circular cylinder than near a plane surface, because the boundary is nearer in other than the shortest direction to it.

The models concerned are cylinders of infinite extent with the flow parallel to the axis, the infinite plane being merely a limiting form of circular cylinder. The shape of the boundary is therefore determined by the cross section of the cylinder. It may be assumed that at a point within the cylinder the mixing length depends on the distance to the wall in all directions on the cross-sectional plane. It would not be appropriate to use a function of simply the average distance, since this would be infinite near a plane boundary. One may make the hypothesis, however, that the mixing length is a linear function of the reciprocal of the average "nearness," in the plane normal to the cylinder, to the cylindrical wall.

To state this more precisely, consider the cross section of any cylinder as illustrated in FIGURE 3,a. In this plane let s be the distance from a point to the wall in the direction θ, this angle being measured from any chosen direction. Then the hypothesis is that the mixing length at this point is

$$l = \kappa\left[\frac{2}{\displaystyle\int_0^{2\pi} \frac{d\theta}{s}} + z_0\right] \tag{20}$$

where κ and z_0 have the same significance as before.

Suppose the wall is a plane at distance z from the point in question, as

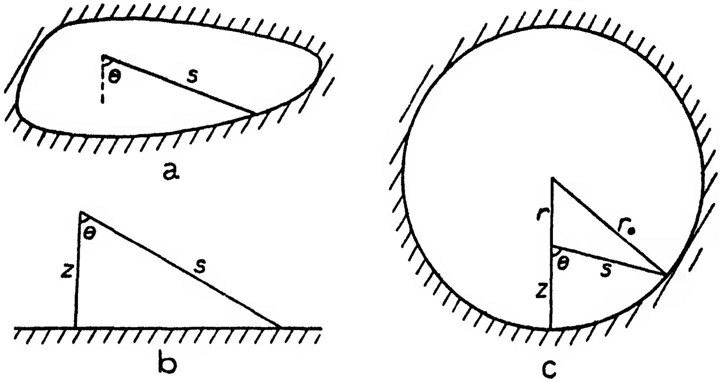


FIGURE 3 Cross sections of cylinders.

shown in FIGURE 3,b. Then the direction may be measured from the perpendicular to the wall, and

$$\int_0^{2\pi} \frac{d\theta}{s} = \frac{1}{z}\int_{-\pi/2}^{\pi/2} \cos\theta \, d\theta = \frac{2}{z}.$$

Thus equation (20) reduces to equation (11), so the assumption is applicable for a plane boundary.

Between two flat plates distant by $2\,r_0'$ the mixing length would be

$$l = \kappa\left[z(1 - z/2r_0') + z_0\right] \qquad (21).$$

At the mid-plane this gives the relative mixing length as

$$\frac{l_m}{r_0'} = \frac{\kappa}{2} \qquad (22),$$

neglecting the roughness length.

The relative mixing length at the axis of a circular cylinder is found immediately to be

$$\frac{l_m}{r_0} = \frac{\kappa}{\pi} \qquad (23).$$

Hence, from the empirical expression for relative mixing length, equation (9), the universal turbulence constant would be

$$\kappa = 0.14\,\pi = 0.44.$$

The empirical determination of relative mixing length near the axis is rather uncertain, however, and the velocity defect will be found better fitted by the value 0.45.

For an arbitrary point within a circular cylinder it is seen by reference to the triangle in FIGURE 3,c that

$$s = \sqrt{r_0^2 - r^2 \sin^2 (\pi - \theta)} + r \cos (\pi - \theta).$$

Hence

$$s/r_0 = \sqrt{1 - k^2 \sin^2 \theta} - k \cos \theta,$$

and

$$\frac{r_0}{s} = \frac{\sqrt{1 - k^2 \sin^2 \theta} + k \cos \theta}{1 - k^2}.$$

Now

$$\int_0^{2\pi} \cos \theta \, d\theta = 0$$

and the elliptic integral

$$\int_0^{2\pi} \sqrt{1 - k^2 \sin^2 \theta} \, d\theta = 4 \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 \theta} \, d\theta,$$

so the relative mixing length is given by

$$\frac{l}{r_0} = \kappa \left[\frac{1 - k^2}{2 \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 \theta} \, d\theta} + \frac{z_0}{r_0} \right] \qquad (24).$$

Neglecting the term containing the roughness length, and using 0.45 as the universal turbulence constant, this expression is shown in FIGURE 1 as curve A. The agreement with observation is seen to be satisfactory.

THEORETICAL EXPRESSION FOR THE VELOCITY DEFECT IN A CIRCULAR CYLINDER

In finding the velocity defect, the roughness length will be neglected. Consequently the shear becomes infinite at the wall and the result will not be applicable close to the wall. Combining equations (10) and (24),

$$\frac{u_m - u}{u_{*0}} \kappa = 2 \int_0^k \int_0^{\pi/2} \frac{\sqrt{k} \sqrt{1 - k^2 \sin^2 \theta}}{1 - k^2} \, d\theta \, dk \qquad (25).$$

The mathematical function on the right side can be evaluated by the use of two expansions in series,

$$1/(1 - k^2) = 1 + k^2 + k^4 + \dots \qquad (k^2 < 1),$$

$$\int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 \theta} \, d\theta$$

$$= \frac{\pi}{2}\left[1 - \left(\frac{1}{2}\right)^2 k^2 - \left(\frac{1 \times 3}{2 \times 4}\right)^2 \frac{k^4}{3} - \left(\frac{1 \times 3 \times 5}{2 \times 4 \times 6}\right)^2 \frac{k^6}{5} - \dots\right] \qquad (k^2 \leq 1),$$

so it becomes

$$\pi \int_0^k \sqrt{k}\left\{1 + \left[1 - \left(\frac{1}{2}\right)^2\right]k^2 + \left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1 \times 3}{2 \times 4}\right)^2 \frac{1}{3}\right]k^4 + \ldots\right\} dk.$$

On integration the result may be written

$$2\int_0^k \int_0^{\pi/2} \frac{\sqrt{k}\,\sqrt{1 - k^2 \sin^2\theta}}{1 - k^2}\, d\theta\, dk = \frac{2}{3}\pi k^{3/2} \sum_{n=0}^{\infty} a_n k^{2n} \qquad (26a),$$

$$a_n = \frac{3}{4n+3}\left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1 \times 3}{2 \times 4}\right)^2 \frac{1}{3} - \ldots - \left(\frac{1 \times 3 \times 5 \times \ldots \times (2n-1)}{2 \times 4 \times 6 \times \ldots \times (2n)}\right)^2 \frac{1}{2n-1}\right] \qquad (26b).$$

Some values for this function are given in TABLE 1.

According to equation (26a) the shear vanishes at the axis, a property lacking in both the previous theories. The ratio $(u_m - u)/u_{*0}$ according to equation (25) is drawn as curve A in FIGURE 2, using 0.45 for the universal turbulence constant. The agreement with the observations is satisfactory, and is definitely better than for curve B or curve C.

TABLE 1. VALUES OF THE FUNCTION

$$2\int_0^k \int_0^{\pi/2} \frac{\sqrt{k}\,\sqrt{1 - k^2 \sin^2\theta}}{1 - k^2}\, d\theta\, dk$$

	Function	k	Function
.00	0.000	.50	0.811
.02	0.006	.55	0.957
.04	0.017		
.05	0.023	.60	1.119
		.65	1.301
.10	0.066		
.15	0.123	.70	1.509
		.75	1.749
.20	0.190		
.25	0.267	.80	2.036
		.85	2.395
.30	0.355		
.35	0.452	.90	2.880
		.93	3.290
.40	0.560		
.45	0.679		

DETERMINATION OF THE UNIVERSAL TURBULENCE CONSTANT

In arriving at a suitable value for the universal turbulence constant, principal consideration is given to the ratio $(u_m - u)/u_{*0}$ at a relative

distance from axis of 0.9. This point is chosen far from the axis, the origin through which the curve passes automatically, and yet far enough from the wall to avoid the complications at the boundary. At this point the observed ratio as given below is 6.47 for smooth pipes, determining in comparison with the function in TABLE 1 a value for the universal turbulence constant of 0.445. For rough pipes the ratio is 6.21, determining the constant as 0.464. The average of these two is about 0.45, which was adopted in drawing curve A in FIGURE 2. This curve agrees with the observations sufficiently well throughout its length, so the value appears satisfactory.

SELECTION OF DATA

The most comprehensive data available on the velocity distribution for flow through long circular cylinders are those of Nikuradse, the fluid being water. His first report (1932) gives the results for smooth pipes, the second report (1933) those for rough pipes. In his experiments the speed of flow was measured just outside the pipe at the exit. Consequently the measured speed "at the wall" is not zero; actually it is of the order of magnitude of half the speed at the axis. The measurements close to the wall, at relative distances from axis of 0.96 and 0.98, must therefore be considered questionable.

When the viscous stress is important at the wall, as is the case with smooth pipes or with rough pipes at low Reynolds numbers, the radius r_0 used in computing theoretically the distribution of mean velocity within the turbulent flow should apparently be the radius of the turbulent core, exclusive of the laminar boundary layer (Montgomery, 1940, p. 12). For smooth pipes Nikuradse gives velocity distributions for 16 runs, for Reynolds numbers from 4000 to 3 240 000. The relative thickness of the laminar boundary layer, computed from

$$\frac{\delta}{r_0} = 7.8 \frac{\nu}{r_0 u_{*0}} \qquad (27),$$

(Montgomery, 1940, p. 14) where δ is the thickness of the laminar layer and ν is the kinematic viscosity, ranges from 0.055 to 0.00014. In order to eliminate the difficulty in regard to the radius of the turbulent core, those runs have been chosen for which the thickness of the laminar layer may be considered negligible compared with a tenth of the radius, the latter being the distance from the wall at which the velocity is important for present purposes. A relative thickness of the laminar layer of 0.002 may be considered negligible, leaving 9 acceptable runs, namely those with a Reynolds number greater than 205 000.

The mean values of $(u_m - u)/u_{*0}$ for these 9 are given in TABLE 2, and are the points plotted in FIGURE 2. The mean values of relative mixing length for these same 9 runs are the points plotted in FIGURE 1 (Nikuradse, 1932, Zahlentafel 5).

TABLE 2. MEANS OF OBSERVATIONS OF $(u_m - u)/u_{*0}$ AS FUNCTION OF RELATIVE DISTANCE FROM AXIS IN CIRCULAR CYLINDERS

$\frac{r}{r_0}$	$\frac{u_m - u}{u_{*0}}$ Smooth[a]	Rough[b]
.00	0.00	0.00
.02	0.02	0.02
.04	0.04	0.06
.10	0.15	0.17
.20	0.42	0.44
.30	0.77	0.79
.40	1.22[c]	1.22
.50	1.79	1.77
.60	2.48	2.44
.70	3.35	3.31
.80	4.53	4.42
.85	5.37	5.16
.90	6.47	6.21
.93	7.46	7.11
.96	8.96	8.45
.98	10.55	9.94
1.00[d]	15.15	14.05

[a]Mean of the 9 runs for smooth pipes with Reynolds numbers 205 000 and greater, from Nikuradse (1932, Zahlentafel 7).
[b]Mean of (1) the 4 runs for rough pipes having relative roughness 1/507 and Reynolds numbers 186 000 and greater and (2) the 3 runs for relative roughness 1/252 and Reynolds numbers 202 000 and greater, from Nikuradse (1933, p. 18).
[c]Nikuradse's value of 1.56 at Reynolds number 3 240 000 is a misprint. The value 1.16 has been used here, as computed from the speeds and friction velocity (Nikuradse, 1932, Zahlentafel 2 & Zahlentafel 4).
[d]Note that these values are computed from the measured speed at unit relative distance from axis, measured outside the exit of the pipe.

With a rough boundary it is customary to reckon distance from the boundary as the distance from the mean elevation of the boundary. In accord with this Nikuradse's values for r_0 are the mean radii of his pipes. Similarly in equations (11) and (20) it may be assumed that z and s are so reckoned. Since there may be some question about this procedure, however, the uncertainty has been eliminated by rejecting the runs in pipes having a large *relative roughness*, the ratio of the diameter of sand grains forming the roughness elements to the radius of the

pipe. The runs were chosen from the two groups having smallest relative roughness, namely, 1/507 and 1/252. From these groups were used again only the runs with a computed relative thickness of the laminar layer of less than 0.002, leaving a total of 7. The means for these are given in TABLE 2 and in FIGURE 2.

DISCUSSION OF THE PAPER

Prof. Kurt O. Friedrichs (*New York University, New York, N. Y.*) communicated by letter previous to the conference:

It is really surprising how well Dr. Montgomery's hypothesis fits with experimental results. It would be quite natural to try similar hypotheses, as for example

$$l = \kappa \left[\frac{1}{\sqrt{\frac{2}{\pi}\int_0^{2\pi} \frac{d\theta}{s^2}}} + \right.$$

instead of equation (20), which is also consistent with relation (11). For a circular cylinder this would lead to the simpler formula

$$l = \kappa \left(\frac{1 - k^2}{2} r_0 + z_0 \right).$$

If this is fitted so as to give $l/r_0 = 0.14$ for $k = 0$, $z_0 = 0$, one obtains for the turbulence constant the value, $\kappa = 0.28$. It is thus apparent that this simpler formula represents the experimental results rather poorly in contrast to Dr. Montgomery's more involved formula.

Of course Dr. Montgomery's hypothesis is of an empirical character. It refers to the situation in the large and thus is at variance with the property of the basic laws of nature to be expressible as relations in the small. The similarity hypothesis of von Kármán, on the other hand, was so devised as to have that property. However, this need not be an objection. It is not unlikely that the stability of unsteady flow depends materially on the situation in the large and that accordingly turbulent flow cannot be characterized by conditions in the small.

Dr. William F. Whitmore (*Naval Ordnance Laboratory, Navy Yard, Washington, D. C.*):

The question of the exact form of average employed in getting an average distance to be used in the expression for l might be given more consideration. An arithmetic average is unsatisfactory, as mentioned, but there are innumerable averages possible in addition to the one used. One which comes to mind immediately is

$$1 \Big/ \sqrt{\int_0^{2\pi} \frac{d\theta}{s^2}}.$$

(Note communicated later.—Subsequent calculation shows the elliptic integral to be replaced by a constant, giving a parabolic law which Dr. Montgomery informs me is in disagreement with the observations, so this average is not so desirable—there is still need for a final decision on the best average, however.)

It is possible, of course, that the resulting curve is not very sensitive to such changes, but quite possibly the value of κ might be altered. Ideally, of course, the form of average used should be determined by general considerations of hydrodynamical theory, but that is perhaps too much to expect at present.

Prof. A. F. Spilhaus (*New York University, New York, N. Y.*):

The average used in the expression for mixing length is taken over a plane angle; might it not be further generalized so as to be taken over a solid angle?

Prof. B. Haurwitz (*Massachusetts Institute of Technology, Cambridge, Mass.*):

Professor Spilhaus remarks that the integral from which the mixing length is obtained is extended only around the circumference of the cylinder and that it would

seem more appropriate to extend the integral over the whole cylinder surface. This remark leads to the question whether the theory can be extended to flow in non-cylindrical vessels or around obstacles.

Reply by R. B. Montgomery:

The problem considered here is strictly one of two-dimensional flow such that the direction of the mean velocity is everywhere parallel to a cylindrical boundary and that its magnitude varies only in a plane normal to the cylinder. The model is of infinite extent, so the turbulence is fully developed. This has the consequence, apparently, that an expression for mixing length can be set up which involves only the geometry of the boundary. The expression contains an average, and this problem can be characterized completely by an average limited to a plane. In suggesting a solid-angle average Prof. Spilhaus implies that the resulting expression for mixing length might apply to other problems, as for example to turbulent flow past a sphere. If some modification of the concept of mixing length should be useful in these other problems, it would not seem reasonable that mixing length would depend only on the geometry of the boundary —mixing length would presumably not be constant on a concentric shell about a sphere. The approach seems inherently limited to the problem specified, and generalization to include other problems does not appear possible.

Prof. B. Haurwitz:

Though the more accurate determination of κ is certainly important and useful, the foremost merit of Dr. Montgomery's paper seems to lie in the fact that it links up the case of flow over a plate, that is, a cylinder of infinite radius, and over a cylinder.

REFERENCES

1. **Kármán, Theodor von**
   1930. Mechanische Ähnlichkeit und Turbulenz. Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse. 58-76.
2. 1934. Turbulence and skin friction. Jour. Aero. Sci. **1**: 1-20.
3. **Lettau, Heinz**
   1939. Atmosphärische Turbulenz. Leipzig 283 p.
4. **Montgomery, R. B.**
   1940. Observations of vertical humidity distribution above the ocean surface and their relation to evaporation. Papers in Phys. Oceanog. and Meteor. pub. by Mass. Inst. Tech. and Woods Hole Oceanog. Inst. **7** (4). 30 p.
5. **Nikuradse, Johann**
   1932. Gesetzmässigkeiten der turbulenten Stromung in glatten Rohren. Forschungsheft **356**, Beilage zu "Forschung auf dem Gebiete des Ingenieurwesens" Ausgabe B Band **3** September/Oktober. 36 p.
6. 1933. Strömungsgesetze in rauhen Rohren. Forschungsheft **361**, Beilage zu "Forschung auf dem Gebiete des Ingenieurwesens" Ausgabe B Band **4** Juli/August. 22 p.
7. **Prandtl, Ludwig**
   1925. Bericht über Untersuchungen zur ausgebildeten Turbulenz. Zeit. f. angew. Math. und Mechanik **5**: 136 139.
8. 1932. Meteorologische Anwendung der Stromungslehre. Beit. z. Phys. der freien Atmos. **19**: 188-202.
9. **Rossby, C.-G., & Montgomery, R. B.**
   1935. The layer of frictional influence in wind and ocean currents. Papers in Phys. Oceanog. and Meteor pub. by Mass. Inst. Tech. and Woods Hole Oceanog. Inst. **3** (3). 101 p.
10. **Rouse, Hunter**
    1938. Fluid mechanics for hydraulic engineers. New York. 422 p.

Annals of the New York Academy of Sciences
Volume XLIV, Art. 2, Pages 105-188
June 8, 1943

CRITERIA FOR VERTEBRATE SUBSPECIES, SPECIES AND GENERA*

By

Charles M. Bogert, W. Frank Blair, Emmett Reid Dunn,
E. Raymond Hall, Carl L Hubbs, Ernst Mayr,
and George Gaylord Simpson

CONTENTS

*This series of papers is the result of a joint symposium held by the American Society of Ichthyologists and Herpetologists and the American Society of Mammalogists at The American Museum of Natural History, April 3, 1942. Manuscript received by the editor, June 1942.
Publication made possible through a grant from the income of the Permanent Fund.

INTRODUCTION

By

Charles M. Bogert

The American Museum of Natural History, New York, N. Y.

Meetings of the American Society of Ichthyologists and Herpetologists and the American Society of Mammalogists were held concurrently at The American Museum of Natural History in 1942. The two societies agreed to a joint meeting for the express purpose of discussing criteria for the lower taxonomic categories, and invitations for formal participation in the symposium were extended by the Local Committees to outstanding workers in the fields of mammalogy, ornithology, herpetology, ichthyology, vertebrate paleontology and vertebrate genetics. The men who accepted were asked to present papers dealing primarily with methods used in vertebrate classification, as the title chosen for the symposium implies.

The modern systematist in actual practice is concerned with the inferences he can draw from morphological characters. But to conform to modern standards he must avail himself of information supplied by other branches of biology. He is no longer interested merely in the preparation of card indexes of names associated with museum specimens. Systematics has become a focal point for many branches of biology, including genetics, ecology, cytology, and psychology, to cite only a few. Synthesis of the information supplied by these various fields of specialization, in part at least, is the task of the modern systematist. He is interested not only in orderly phylogenetic arrangements of organisms, both fossil and recent, but in evolutionary processes. As Julian Huxley has recently stated, "The problem of systematics is that of detecting evolution at work." No taxonomist can afford to be disinterested in such pertinent biological phenomena as isolating mechanisms, nor will he discount the value of population studies in genetics. Contributors to the symposium call attention to the value of theories founded on data supplied by other fields, but applicable to taxonomic concepts. Improvements in definition are not lacking as a result.

The viewpoint expressed by each of the contributors is not necessarily the consensus of workers in the fields represented. To a large extent it is a personal viewpoint, dependent upon the training and experience of the individual, but in part it has evolved from the nature of the material

with which each has been chiefly concerned. Each author places stress upon one factor or another in defining the criteria he advocates. Still it is obvious that each has been confronted with similar problems.

Criteria for genera cannot be defined on any grounds and, needless to say, no rule of thumb has been attempted, if the generic category has been discussed at all. Nevertheless some very worthwhile considerations have been set forth, particularly in the papers by Simpson and Mayr.

The importance of the subspecies in the taxonomic system is not questioned, but several difficulties in defining criteria for infraspecific groups are discussed. The differences between subspecies, races, subraces, colonies, and local populations are not clear cut; indeed the terms have no formal, precise connotation in most vertebrate work. "Race" and "subspecies" seem to be used as synonymous terms except in the field of ichthyology. Further complications exist in interpreting inferences drawn from data where gradients or clines exist and the ideal of objectivity is perhaps not attainable.

The species is recognized as the fundamental unit of taxonomy, but all contributors do not concur with reference to criteria for the species. The strongest dissenting voice emanates from the field of mammalian genetics. Blair regards intersterility as the criterion for species. Others place more stress upon conditions in nature. Potential hybridization under laboratory conditions is not considered of great importance in determining taxonomic rank. Cases are cited where, on the basis of inferences drawn from morphological characters, two populations apparently interbreed in one area where ranges overlap, but fail to produce intermediates or hybrids in other areas where representatives of both occur side by side. Whether the problem is essentially ecological or genetical it concerns the taxonomist.

Problems of inference are inherent in much of the discussion and each contributor freely admits the arbitrary nature of his criteria. For the most part contributors are in essential agreement whether they define species in terms of "absence of intergradation," of "reproductive isolation," or mention "closed systems," "discontinuous units" or "differentiated kinds." It may be said that the major disagreements lie in the conjectural interpretation of similar data. Controversial issues raised in one paper are frequently answered in another. In this respect, and in the diverse viewpoints expressed, the individual papers are complementary. The symposium, as herewith published, represents not a mere description of taxonomic procedure but a summary of information relevant to the problems involved in the interpretation of the processes of evolution.

CRITERIA FOR SUBSPECIES, SPECIES AND GENERA, AS DETERMINED BY RESEARCHES ON FISHES

By

Carl L. Hubbs

University of Michigan, Ann Arbor, Michigan

The criteria for the minor taxonomic groups of fishes do not differ essentially from those that are pertinent to the subspecies, species and genera of the tetrapod classes. The increasingly detailed and penetrating researches of recent years keep emphasizing the conclusion that these categories are of essentially the same sorts in all groups of vertebrates. Speciational processes appear to be similar throughout the Vertebrata, in fact throughout most of the organic world.

For several reasons, however, we find that fishes are particularly well suited to an analysis of the minor taxonomic categories, and hence to an appreciation of the criteria by which these assemblages can be recognized. In the first place, fishes are rich in differentiae: they fairly bristle with diagnostic external features, and characters penetrate their whole anatomy (the internal characters may readily be investigated, for the whole bodies—not just skins—are preserved). Most of the distinguishing features are readily subject to precise evaluation, and hence to statistical analysis; meristic differences are common. For such studies it is often possible to make use of hundreds or even thousands of specimens, for large series of fish are readily collected and preserved. The critical relations between individual and racial modifications are particularly amenable to study in fishes, for the aquatic vertebrates are molded by their environment to a greater degree than are the representatives of the so-called higher groups; the trend of vertebrate evolution has been to free the organism from domination by its physical environment. The fact that many of the stocks or races within a fish species occupy highly diverse ecological situations often makes it possible, without recourse to experiment, to analyze the individual modifications and the genetic responses to the environment. Finally, the genetic basis of characters and the genetic interrelationships of forms can be determined more readily by experiment in the Pisces than in most other groups, because many fishes may readily be propagated, and because of the frequent occurrence, in certain fishes, not only of subspecific intergradation,

but also of interspecific and even of intergeneric hybridization. Natural hybrids between species seem to occur more commonly in fishes than in any other group of animals. Fishes are especially well suited to the study of forms as populations of living animals, not merely as samples of preserved specimens.

Except as they may be arbitrarily erected, the distinguishing criteria cannot be more precise than the true nature of the taxonomic categories. Every advance in the general systematics of fishes tends to dispel further any idea that these categories differ from one another with sufficient uniformity to permit lines of clear demarcation to be drawn consistently between them; or that any of the assemblages are subject to a very precise and exclusive definition. He who creates hard and fast distinctions between the systematic categories is merely drawing lines of chalk down a blackboard.

Largely on the basis of convention, we recognize a primary distinction between kinds or forms (colonies, subraces, races, subspecies, and species) on the one hand, and groups (species groups, subgenera, genera, and larger categories) on the other side. The kinds are regarded as the units which make up the groups, but each kind may ordinarily be shown to be a group comprising lesser units, down to the small colony and finally to the individual (which in turn is an integrated colony). And each minor superspecific group, such as a genus, may be scarcely less a kind than is the species; this is most obviously true of monotypic genera. It is therefore difficult to state criteria, even for the distinction of kinds from groups. All kinds of animals and all of the groups are surely to be regarded as graded levels of evolutionary differentiation.

Races, subspecies and species are all regarded as kinds or forms, and therefore as having common terms in their definition. Kinds may be defined as self-perpetuating populations with a considerable degree of uniformity in time and in space. There are then three essential considerations in the concept of animal kinds: genetic basis; integrity as populations; success in the struggle for existence. Somatic responses of the individual are not true forms. Systematic characters must have a genetic basis; if this is not demonstrable, it is to be inferred. But one or a few genetically distinct animals do not necessarily constitute a systematic kind. There must be a population, self perpetuating and distinctive. Animal kinds are not made up of characters, or of museum specimens, nor of age or sex variants, nor of interbreeding phases; they are living entities in a living world. And, as David Starr Jordan has insisted, populations with a genetic basis do not qualify as species (or as infraspecific units) until they have successfully run the gauntlet in the

vigorous struggle for existence in nature. In speciation, survival is quite as significant as genetic potential.

The fish data seem particularly valuable in the classification of the types of animal kinds. In addition to the usual distinction of species and subspecies, the race is interpreted as a distinct taxonomic category by most ichthyologists. Unlike many ornithologists and mammalogists, they do not use "race" or "geographic race" as a synonym of subspecies. For several reasons, such as the abundance of available material and the demand for the population-analysis of commercial fishes, minor races have been studied more in the Pisces than in other groups. Ichthyologists have therefore had most reason to separate the race as a category distinct from subspecies. They think the distinction to be a valuable one and urge its general adoption. Such studies as those of Sumner and of Dice on *Peromyscus* and of Miller on *Junco* suggest that the concept of the race, as a lesser division than the subspecies, will come into wide use through the vertebrates.

Races, in the sense here advocated, are not accorded a place in the current system of zoological nomenclature. One reason is that since such races are ordinarily distinguishable only by average characters that may call for statistical treatment, the routine identification of single specimens or small samples would often be difficult or even impossible. Race ranking may be accorded forms, like local types of *Gasterosteus aculeatus*, which are so confusingly numerous or so complex in characters, and so complicated in genetic and geographical relationship, as to transcend any ordinary scheme of zoological nomenclature. Excluding the races from the nomenclatorial system does not imply that their distinction and study are unimportant. On the contrary, the minor races--potential species in the early stage of their making—are perhaps the most significant material for the study of speciation.

For one, I decry, as impracticable and as contrary to the International Rules, the use of a quadrinomial system for the designation of races, and I still more vehemently oppose the nomenclatorial recognition of "morphae" and "nationes," as distinct from subspecies.

Unless the systematics are excessively complicated, I would designate as a subspecies any genetic form which shows reasonable geographical or ecological consistency, and which can usually be distinguished on its totality of characters. Ordinarily it would be required that much more than half of the given population be distinguishable; not necessarily at all times and places, but at least in one sex, at some given stage of development. Interpreting the ensemble of characters, as by the method

of the character index, may aid in defining a form as a subspecies rather than as an unnamed race.

Unlike races, subspecies are animal kinds which are sufficiently clear-cut as to be thought worthy of a place in the nomenclatorial system, but which do not give evidence of being completely differentiated. Species, in simple terms, may be denoted as kinds (defined above) which are completely differentiated from all others, but which do not themselves include fully differentiated subdivisions. These definitions of the categories of animal kinds are based on practical as well as theoretical considerations. Recent studies in ichthyology seem to justify the compromise.

Despite the obvious intergradation between systematic categories in fishes, ichthyologists share with other systematists the view that a higher degree of distinctiveness and reality is possessed by the species, than by any infraspecific or any supraspecific category. However diverse our opinion may be as to the magnitude or consistency of this distinction, or however much we may confuse the issue by philosophical quibbling over "reality" versus "concept," we agree in the view that the species represents a rather definite and relatively stable level of attainment in evolution.

The creation of a new species marks the approximate end of the speciational process, for differentiation has then proceeded to completion or nearly so. Species, therefore, are usually trenchantly distinguished from one another. The ichthyological evidence leads us to suspect that species dominate the systematic picture, not only because of their distinctness but also because of their long life. Once the specific level has been attained, with its more or less complete genetic segregation, further divergence will probably be accelerated, but the local and individual variants that appear will ordinarily long remain an interbreeding part of the species complex. The further transformation or division of the species into one or more new species is a hurdle which Nature finds difficult and time-consuming to jump.

In my younger, more radical, years I was so much impressed with the view that species are the most natural and distinctive of the taxonomic categories, that I considered proposing a uninomial system of zoological nomenclature. Such a system, it was thought, would have many advantages over the Linnaean one which has been in use for nearly two centuries; it would indeed divorce taxonomy from nomenclature, lead to the stability, uniformity and brevity of names, and force precision in identifications; and it would be the logical outcome of the tendency

toward monotypic genera. If subspecies were always sharply distinct from species, such a scheme would indeed have a strong call for adoption.

However, the more intensely I have studied fish species, the greater the number of species I have examined, and particularly the more thoroughly I have studied whole species groups throughout the range of the complex, the more difficulty I have found in the interpretation of forms as subspecies or as species and in distinguishing species from species groups. The distinctions appear more and more arbitrary. This realization, along with the attainment of a more conservative age, has led me to abandon the idea of proposing a new system of nomenclature.

Possibly the fresh-water fishes (of which I have studied more than a million specimens) are peculiar in the high percentage of forms which only by arbitrary decision can be classed as subspecies or as species. On the other hand, marine fishes and other groups of vertebrates may seem to exhibit sharp distinctions between species and other taxonomic categories, only because the systematic studies have been based on fewer specimens or have been less analytical, or because the systematic picture has been more dominated by convention and "authority." I suspect that the truth involves both alternatives.

Incompleteness versus completeness of differentiation is the main test by which subspecies may be distinguished from species on the kinetic concept here advocated. Degree of differentiation is difficult to determine, especially from small samples, but is the truest measure we can obtain of the stage of speciation. No other criterion would seem to have so sound a speciational basis, or to be so consistently applicable.

Determining or predicting whether two kinds do or do not intergrade therefore becomes a matter of paramount nomenclatorial importance. Intergradation is a central problem in practical systematics as well as in speciational research.

Intergradation would be rather easy to demonstrate, if it were regularly of the simple type which alone has been recognized by many vertebrate zoologists. According to their oversimplified view, intergrading forms have well-defined ranges which are separated by a narrow belt in which the characters of the one subspecies grade rapidly and evenly into those of the other. Recent variational analyses by ichthyologists, however, keep emphasizing the views: (1) that this simple, narrow-band type of intergradation is perhaps the exception; (2) that there are many, often complex types of intergradation; and (3) that, in the area of intermixture, there is every stage between complete fusion and almost complete genetic isolation. Such critical studies as those of Blair on *Bufo*, of

Fitch on *Thamnophis*, of Miller on *Junco*, and of Dice on *Peromyscus*, confirm the view that varying patterns and degrees of intergradation also characterize the higher vertebrates.

The simple intermediate-band type of transition between well-defined subspecies of fish does undoubtedly occur, especially in some forms with coast-wise distribution (for instance in certain West Coast blennies). At times, the area of intergradation may be very broad, greater than the range of either subspecies. Or the area may be small, or may be a longitudinal connecting strip rather than a transverse band. Commonly there is a chain of forms, or a rather even gradation from one end of the range of the species to the other. Among the fresh-water fishes intergradation tends to follow what we have termed a mosaic pattern.

The ranges of two subspecies are commonly separated by a wide belt, in any part of which we may find: (1) entire populations typical of either form; (2) mixtures of the two types; (3) a complex of either one or both types plus intergrades; or (4) intergrades alone (*Notropis cornutus chrysocephalus* and *N. c. frontalis*, for example, intergrade in all these ways). If we free ourselves of preconceived ideas of how subspecies are or should be interconnected, we can readily appreciate why such complex patterns are probably the rule. Two forms which occupy distinct territories, with different climatic conditions, will almost surely develop different environmental responses. For this reason the two types, in the intervening zone, will tend to segregate themselves, or, if they occur together, they will often spawn in different niches or at separate times. Where the environmental conditions are contrasting, the two types will tend to maintain their identities; where the environment is uniform, the forms will intergrade.

Such diverse types of intergradation will probably be found to hold throughout the vertebrates. To treat as subspecies only those forms which intergrade in the conventional pattern would be indefensible.

The common tendency, at least in fresh-water fishes, for two forms to intergrade in some localities but not in others, renders difficult and equivocable the test that is most commonly applied, usually without doubts or reserve, to determine whether two forms should be separated as subspecies or as species. This is the test of determining whether two forms that live together do or do not intergrade. Further difficulties in applying this test arise from the incompleteness of available data, and from the subjectiveness that must ordinarily be involved in reaching an "inference" concerning the genetic relationship; from the circumstance that the really critical data pertain to the exact time and locale of breeding; and from the circumstance that two forms may occur together with-

out interbreeding, yet be connected by a chain of interconnected subspecies (the "open circle").

Then, too, we encounter the uncertainty, that forms morphologically intermediate between two kinds may arise through the interbreeding of these two kinds. The interjacency of characters may be due to chance variation, or to independent adaptations to intermediate conditions. On the subspecies concept that requires intergradation as a result of interbreeding, such interpretations would throw the systematic decision into doubt. In my view, however, the existence of truly intermediate types would be taken as evidence that subspecies are involved, even though the intermediates had not arisen through interbreeding. I would not even demand that the intermediates occupy an intermediate range: they may crop out at any interior, peripheral or even disconnected location.

In general I subscribe to Jordan's Law, that nearest relatives tend to occur in adjacent, rather than in the same or in distant territories. I therefore regard geographically contiguous but complementary ranges as typical for subspecies. It would, however, be unwise to accept this relation as a criterion for subspecies. Forms otherwise like subspecies, and regarded by me as such, may occur in widely separated regions, or on narrowly but completely separated ranges (as in adjacent stream systems tributary to the ocean). Again we find forms of the subspecies type whose ranges are entirely enclosed within that of their cognates: thus *Boleosoma nigrum eulepis* exists in appropriate pockets within the range of *B. n. nigrum*, with a halo of intergrades around each of the units of discontinuous distribution.

In increasing number, subspecies in fishes are being shown to be ecological (or microgeographical) forms, which occupy diverse habitats in the same or in very broadly overlapping areas. They exhibit a pattern of partial or complete intergradation, in which the intergrades occur mosaically at appropriate points throughout the common range of both forms. This type of subspeciation is proving to be common in fishes, notably among the fresh-water fishes of the western United States. *Notropis volucellus* is an outstanding example in eastern North America. Most geographical differentiation is in part ecological, and most ecological subspecies have some geographical basis. The two types completely intergrade. Obviously, then, the criterion of intergradation is not to be restricted to purely geographical interconnections.

Not only may subspecies violate the rule of adjacent, complementary ranges, but full species may show the distributional type that is more typical of subspecies. Many unquestionable species—even some genera

—of fish show complementary distribution. It could hardly be expected that geographically seriated subspecies would never differentiate *in situ* into seriated species.

Character intergradation between two wholly or almost wholly isolated (non-interbreeding) populations is not, in my opinion, to be excluded as evidence of incomplete differentiation, and may therefore be taken as a criterion of the subspecific status of the two kinds. It is probable that some of the intergradation that is observed between forms of overlapping or contiguous ranges is due to parallel adaptations, of separate origin, to the intermediate conditions, or to chance variation rather than (or in addition to) interbreeding. The criterion for subspecies, of intergradation due to cross-mating only, would seem to be impracticable as well as illogical.

Thus I regard intergradation of almost any type as evidence that speciation is not complete and that the forms involved are on the subspecies level of differentiation. I would, however, lay much greater stress on actual intergradation in nature than on potential interbreeding. I do so not only because I lean quite as heavily on the survival as on the genetic aspect of the concept of animal kinds, but also because it is more objective to deal with what exists than with what we may infer to be possible. The curse of most systematic work has been its subjective or "authoritative" basis, against which we should react. What actually occurs seems to be most significant, and it is generally admitted that many apparently "good" species of fish and other animals interbreed freely in captivity yet never or very seldom do so where they occur together in nature. There are other than purely genetic isolating mechanisms. As Haldane has written, "The physiological barrier between two species may be of several different kinds. It may occur before or after fertilization, and the hybrids, if any, may be sterile, or more or less completely fertile." I prefer to regard as full species any two completely distinct forms which do not intergrade in nature, whether or not they fail to interbreed because of their isolated ranges. If under appropriate circumstance the two forms at any time come together and interbreed regularly, I would take this as evidence that they have reverted from the specific to the subspecific category of differentiation.

Thus I regard it as impracticable to restrict the concept of subspecific intergradation to connections between forms that are assumed to be diverging; that is, to what may be called primary or antecedent intergradation. It will commonly be difficult or impossible to determine whether the intermediate characters of a given population date back to the initial divergence of the two types, or are due to the secondary or

subsequent meeting and interbreeding of populations, or of entire forms, which had previously been isolated for some time. Our analyses of intergrading forms lead us to believe that most intergradation, even in a single region, is due to a combination of the primary and the secondary processes. It would seem wholly impracticable to set up divergence, as opposed to convergence or fusion, as a criterion of the subspecific level.

It would also be unwise to restrict the concept of subspecies to forms which change as organismic units from one to another across their zone of intergradation. An increasing number of cases are being discovered, in fishes as well as in other groups, in which the units of geographical variation and intergradation are characters which do not all change at the same region. Thus, as we proceed southward, a color feature (for example) may intergrade at one point, whereas a scale character will change farther south. If the intervening type with the color of the southern race but the squamation of the northern one has a considerable integrity in characters and in range, it may be recognized as an intermediate subspecies. The same treatment may be accorded intermediate populations which exhibit harmonious or concurrent intergradation in all characters. Whether or not to recognize the intervening subspecies should depend on individual and geographic consistency in characters.

Furthermore, the two sexes may intergrade at different points.

Commonly the specific level is evidenced by completeness of differentiation in each of several respects, usually in at least one character. Complete differentiation, however, may be demonstrated only when the ensemble of characters is studied. Each differential feature when considered alone may show an overlap, when the frequencies for the two types are compared. Nevertheless, the kinds may invariably be distinguished on the basis of the ensemble of their characters. This completeness or near-completeness of differentiation may often be statistically demonstrated by the use of the character-index method which we have been employing.

We do not often hear, in modern ichthyology, of the formerly common expression, "specific characters," as something essentially different from "varietal characters" or from 'generic characters." Consistency is a better test than kind for the taxonomic significance of a character. Thus we find that the union of the lower pharyngeal bones coupled with the transformation of conic pharyngeal teeth into molars varies considerably within a species, and in different groups of fishes characterizes a subgenus, a genus, a family, and an order—according to the current system. Characters, like gold, are where you find them.

Despite certain claims, the mode of inheritance of the systematic

characters provides no valid criterion for systematic ranking. We can not place a given form as a subspecies or as a species, on the basis of whether its characters do or do not show sharp Mendelian segregation. Our breeding experiments with fishes and our analyses of natural hybrids are indicating that most of the many systematic characters studied, whether of race, species, or genera, behave according to a seemingly Galtonian type of inheritance, though some features of equal systematic value do Mendelize simply. This would seem to be the general rule, at least for vertebrates. Haldane wrote in 1938: "The majority of inter-specific differences, however, blend, though there is usually an increased variability in F_2 which can be explained as due to segregation . . . The hypothesis of multiple factors is at present neither proved nor disproved."

In general, however, we find that simple Mendelian segregation is typical of "sports" and "phases" that do not occur with sufficient consistency as discrete populations to warrant their inclusion in the taxonomic system. The hereditary pigment types of *Lebistes*, as studied by Winge and others, and the caudal-peduncle markings of *Platypoecilus*, so thoroughly investigated by Gordon, I would put in the same class.

Some have held that subspecies produce intergrades uniformly intermediate in all characters, whereas species, when they interbreed, yield offspring with a mixture of the characters of the two parental types. The extensive evidence on intergrades and hybrids in fishes does not support this claim

Evidence from reciprocal crosses between subspecies and species of fishes contradicts the odd view that "nuclear differences may account for variation within a species," whereas "the deeper differences between species depend on the cytoplasm."

In a general way only it may be said that subspecies exhibit adaptive responses to temperature and other physical features of the environment, whereas species characters, if adaptive, are correlated with food habits and other biotic factors. Such relations, however, do not hold with any great consistency and are not available as definitive criteria for systematic ranking.

There may still be some vertebrate zoologists—few ichthyologists, I hope—who would accept as a criterion for systematic ranking, the interpretation of the characteristics of the animal as adaptive or non-adaptive. Some have held that subspecific characters are adaptive, whereas specific and generic features are not of survival value; others deny that species possess the adaptive characters which the superspecific groups display. Such a criterion would be unsatisfactory, in that it would tend to maintain systematics as a subjective rather than an ob-

jective art. I would add that adaptiveness is not rightly to be used as a criterion for systematic ranking, because, in my opinion, almost all characters—whether of races, subspecies, species, genera, or higher groups—have been involved, primarily or secondarily, in the adjustment of the animal to its particular environment.

To rank two forms as subspecies because "their intimate relationship would be concealed by a grant of specific status" seems to be an unjustified taxonomic procedure. Of course subspecies are more closely related than species, but this is true only on the average. To class a form as a subspecies by reason of one's opinion as to relationship carries all the dangers that commonly go with subjective decision. The next worker may have a different feeling. Stability of nomenclature calls for more objective criteria.

It is similarly wrong, I would say, to force a kind into the subspecies rank merely because it is a member of a minor evolutionary cluster of which the other units are subspecies. There is no reason why any member of such a group should not attain complete differentiation and hence warrant specific status.

Ordinarily an inferred monophyletic origin is regarded as a requirement for any natural group or kind. Lately some authors, notably Dice, have interpreted subspecies as ecological responses which may originate repeatedly. Such local types as the black mammals of the lava-beds, or the brackish-water races of *Zoarces*, very probably have evolved independently in response to a like habitat. Monophylety is obviously not to be applied rigidly to infraspecific units. Until complete genetic isolation has been attained, forms will no doubt often interchange their genes. Identical or similar mutations, followed by selection, could readily transform a given kind of animal repeatedly into adaptive products that are indistinguishable or even genetically identical. The polyphyletic races of a subspecies may differentiate into distinct species; or a character which will later define a distinct species or even genus may arise independently at several localities. Hence, to some degree species and genera, as well as subspecies, may be polyphyletic.

One of the better distinctions between subspecies and species lies in the magnitude of the structural differences. But again no clear-cut, usable criterion is provided. Even within a race, terata and phases may show differences greater than those separating related genera. Some subspecies which are evenly connected by intergrades show differences which are more trenchant than many of the specific distinctions in the same group (such, for example, are the differences between *Gambusia affinis affinis* and *G. a. holbrookii*). It is to be expected that bonds of

intergradation may be continuously retained by some forms, which have undergone an amount of differentiation that would ordinarily accompany complete speciation; or that two forms long separated and well differentiated may retain potential interfertility, so that they will again intergrade and become subspecies when their ranges come to overlap. Amount of difference is therefore not an infallible criterion for systematic ranking.

Ichthyologists will concur in the general view that genetic isolation is probably the best single index by which species may be distinguished from subspecies Recent researches on fishes, however, prove that even this test often breaks down. Degree of fertility is positively correlated with degree of relationship, but only in a rough way. Subspecific intergrades tend to be fertile, but some crosses within a single race are sterile; except in certain groups like the Cyprinidae, most species are more or less completely intersterile, or produce infertile offspring, but some genera yield fertile progeny. Furthermore, in some groups there is every gradation in fertility, without the sharp distinction between interfertile subspecies and intersterile species that some geneticists, as Shull, postulate. We find this true, for example, in the genus *Mollienisia*, on which we have been conducting breeding experiments for ten years. Effective isolation in nature often precedes the attainment of sterility, as already mentioned in treating potential as contrasted with actual intergradation. Just where and how to draw the line between genetically connected and genetically isolated forms is often very difficult to determine in fishes, not only those which are little studied but also those which have long been subjected to breeding experiments.

CONCLUSIONS

To my knowledge no single criterion that has ever been erected will suffice to define the species, without the need for some exceptions and modifications. The more intensively species are studied, throughout their ranges, the more difficult it often becomes to decide on the taxonomic rankings. Nevertheless, most species in most groups seem sufficiently distinct to be interpreted as such by all systematists. Even when all single tests for the species level break down, a form may be recognized as a species by reason of the usual validity of a series of criteria, just as some subspecies and species may be known by the usual though not invariable possession of each of a series of characters.

Much more often than is generally supposed or admitted, the distinction of subspecies from species appears indefinite and arbitrary. Among

the general biologists Haldane has appreciated this circumstance, for, in 1938, he wrote that: "Our general conclusion is that there is no evidence that at any rate closely related species differ in a manner qualitatively diverse from varieties." Subspecies no doubt are usually terminal twigs, but, contrary to the views of Bateson and Goldschmidt, some of these twigs no doubt grow into great limbs on the tree of evolution.

We have been surprised to find, of late, that we must be arbitrary even in ranking some forms as subspecies or as genera. This is true of two minnows of the Mohave Desert, which represent two of the wide-spread genera of the West. In the Pluvial period these fishes were obviously separated as lacustrine and fluviatile types but they are now forced into cohabitation in the dwindled desert waters and have hybridized so prolifically that 8 per cent of the total minnow population of the Mohave River system is now composed of intergeneric hybrids, or of subspecific intergrades—depending on judgment. Probably on somewhat similar grounds the blind cave characin of Mexico, *Anoptichthys jordani*, has interbred with its ancestral type, *Astyanax fasciatus mexicanus*, to produce a complete series of what may with equal propriety be called intergeneric hybrids or subspecific intergrades. Our extensive researches into both phenomena have failed to disclose any essential or consistent distinction between subspecific intergradation and interspecific hybridization.

It is perhaps unfortunate for an orderly taxonomy, or for a pretty scheme of speciation, that the systematic situation is so complex. But neither in detailed taxonomic treatment nor in general speciational theory should we forget the true situation. Arbitrary decisions must often be made, to meet the demands of the Linnaean system of zoological nomenclature, but it is bad science to deny that the decisions are arbitrary. Neither conventionalized views nor subjective subterfuges—whether by the old-line systematist or by the modern speciationist—can transcend the facts, or create a simple "correct" system of taxonomy or a simple theory of speciation out of a situation that is inherently complex. Evolution has been and remains at work.

There appear to be no objective criteria for genera.

LOWER CATEGORIES IN HERPETOLOGY

By

EMMETT REID DUNN
Haverford College, Haverford, Pennsylvania

The consideration of lower categories in herpetology may be of general interest because of peculiarities of some of the groups. These are some life histories which differ considerably between closely related forms, and which are observable and open to experimentation from beginning to end; relative ease of field observation and collection; relative ease of maintenance in the laboratory.

As a herpetologist I agree with, and have little to add to, the very considerable body of general considerations on lower categories which has recently been put forth in print (and to which I myself have contributed). Thus, I am in agreement with the following notions, which I shall not at this time labor to support: genera are matters of opinion, personal arrangements of species; species are "distinct self-perpetuating units," "*distincta propagatio ex semine*," discontinuous from other species; subspecies (or races) are populations within a species, as many as may be considered easily recognizable; varieties are minority elements in a population; individuals are the elements of which varieties, subspecies and species are composed.

The characters, presumably genetic, which differentiate these categories morphologically (and physiologically) are all of the same status. There is no distinction between individual, varietal, subspecific, specific, and generic characters. Relative constancy of characters differs widely from group to group but, reflecting, as it presumably does, the genetic makeup of any group, a constant character for a group is inevitably used as a diagnostic character

We postulate some form or forms of isolating mechanism as a historical factor in attempting to account for races (incomplete isolation) and species (complete isolation). No such factor is postulated for varieties. We find the most common form to be spatial (geographical)—sometimes mere distance, more often distance plus difference in nature of surroundings (organic as well as inorganic), very often distance plus difference plus some sort of barrier. Ecological, temporal, and physical isolating mechanisms have also been suggested for some cases in herpetology, and physiological, psychological, and genetic isolating mechanisms may well exist.

Study of herpetology leads me to the generally accepted opinion that descent with modification has taken place in a rather uniform manner, and that the lower categories have come to differ by essentially the same method: gradual replacement of one genetic characteristic by another. No evidence of a distinction between microevolution and macroevolution is evident. The supposed distinction seems to me to rest largely on certain characters being considered as of a higher status than others, a view which I do not share. I would agree, however, that evolution by geographic isolation, not involving much change in the selectional value of hereditary characters, is very common, and might be termed "microevolution," whereas evolution by ecologic isolation, involving considerable change in the selectional value of hereditary characters is rather rare, and might be termed "macroevolution." Both sorts are known to me "*in statu nascendi*" from herpetological examples, and I find no significant differences between the two.

I shall consider species first, then varieties, then subspecies, and lastly genera.

SPECIES

There has recently been a suggestion that the old "*distincta propagatio ex semine*" criterion for a species, "a distinct self-perpetuating unit," a criterion which is perhaps subjective but which is certainly based on field observation, should be replaced by a more objective, genetic, and experimental criterion, a criterion which lays less stress (if any) on field observation.

This new criterion of specific status, incapacity of two forms to breed together or mutual infertility, while logical and objective, may be criticized by the practical taxonomist (a) because it is usually impossible to apply it to the material with which he has to deal; (b) because in many animals actual discontinuity of breeding may occur long before mutual infertility sets in and the criterion if strictly applied would mean that many naturally "distinct self-perpetuating units" would be lumped by it; (c) because cases are known in which two populations of what is demonstrably the same species may manifest actual discontinuity of breeding and act toward each other as two species.

The axolotl (*Siredon mexicanum*) and the *velasci* race of *Ambystoma tigrinum* show no intergradation in nature; they live or did live in the same body of water; in nature they are distinct species, but they are not mutually infertile

In the French Broad Valley and in a valley in the Smokies, *Desmognathus fuscus* intergrades with *D. carolinensis*. The intergrading rela-

tionship between *carolinensis* and *ochrophaeus* is so close that opinions differ as to the boundary between the two. In New York State *fuscus* and *ochrophaeus* are for practical purposes two distinct species. This is a chain of only three links; the middle link intergrades with each of the two ends, but the two ends overlap each other without any indication of intergradation or interbreeding.

VARIETIES

The distinctness of self-perpetuating units is partly a matter for morphological observation, the presence or absence of intermediate characters, the presence or absence of intermediate individuals. A considerable number of cases cannot be settled on grounds of morphology alone, the phenomenon of non-continuous variation necessitating biological observation. (Theoretically it may be maintained that all variation is discontinuous, but practically the distinction between continuous and discontinuous variation is eminently sound.) Most herpetological species exist in two varieties, morphologically distinct, and between which there are no intergrades. I allude to the two sexes. These two varieties, distinct as they usually are, can be relegated to the same species on grounds of biological observation. In some cases we know something of the genetic basis by which these two varieties are maintained in approximately equal numbers in the population.

In 1818 Green described two species of salamanders of the genus *Plethodon* from New Jersey, naming a dark form *cinereus* and a redbacked form *erythronotus*. On morphology alone his action is still justifiable. There is an absence of intergradation (discontinuity), and while the two usually coexist in approximately equal numbers, such is by no means always the case. It is true that Tschudi in 1838 expressed the opinion that the two were conspecific, and that most herpetologists followed his lead, but the fact that the two belonged to the same self-perpetuating unit was first proven by biological observation by Burger in 1935.

As in this instance, most cases of discontinuous variation involving only a single pair of contrasting characters have been dealt with by assumption. Cases in which more than one pair is involved may occur. Thus the North American snake, *Chironius fuscus*, either occurs in four varieties or else there are four species. There is no intergradation and no correlation of variation with age or sex. The snake may be black or spotted, the anal plate may be single or double. All possible combinations of these variations occur. This snake is utterly unlike anything else in North America. I solve this problem by considering the four morphological types as varieties, but until biological proof is produced I

cannot defend my solution against anyone who chooses to consider them four species.

Smith has recently described a Central American snake of the genus *Dendrophidion* in which two quite distinct counts of tail vertebrae seem correlated, one with a single, and the other with a double, anal. He considers them distinct species, and on purely morphological evidence he is right. They are distinct. But there is no biological evidence at all. The two have the same range and occur together, and could be considered as conspecific, linkage of two genes, or two effects of a single gene change resulting in two distinct varieties within the same species.

Cases such as these are of the highest general interest and merit genetic study, as well as more field observation. Surely the means whereby the two different forms of *Plethodon cinereus* are maintained in nature in approximately equal numbers could be investigated.

SUBSPECIES

Within the theoretical (often actual) continuity of breeding individuals of the species, local differences may occur in the incidence of certain characteristics. These differences may differentiate two or more races (subspecies) within the species. There are two criteria generally applied to such cases. One is intergradation. It is characteristic of subspecies that some individuals cannot be properly allocated because: (a) they have intermediate characters; or (b) exist outside the main area of their type. The second criterion is the amount of the local incidence of the traits. It is generally held that at least 75 per cent of the individuals in a given area must be distinguishable from those in other areas of a species range to make it worth while to recognize a subspecies. The majority of such situations are spatial (geographical). A few diagrams may show the most common cases. The delimited rectangles of the diagrams may be thought of as absolutely separated islands; as land areas separated by some more or less impassable barrier; as land areas in complete continuity one with another. The letter A and the letter B each represent a kind of animal, the former differing from the latter by one or more characters. The combination AB represents animals which, in themselves, combine some of the traits of A and some of the traits of B, or have traits intermediate between those of the two extremes. The numerical figures are wholly hypothetical, for purposes of illustration, but quite similar figures could be produced for actual cases.

1. **Intermediate individuals existent; variation continuous; or variation discontinuous and uncorrelated (shuffling).**

a.

A	A	AB	B	B
100	5	90	5	100

b.

A	AB 10
100	B 90

c.

A 90	AB 10
AB 10	B 90

d.

A 90	A 5
AB 5	AB 5
B 5	B 90

2. No intermediate individuals. variation discontinuous

a.

A 100	A 50	B 100
	B 50	

b.

A 100	A 10	A 90	A 10
	B 90	B 10	B 90

I have diagrammed seven possible cases of a species with two vicarious races. Diagram 1,a represents the most common situation. I do not suppose that I have exhausted the possible situations but I have included more than are usually recognized, many systematists having refused to consider any but 1,a.

A species may include many more than two vicarious races. In such cases the line of races may curve and the two ends meet. Intergradation between the two ends may occur as the term "Formenkreis" implies. Just as frequently, the two ends may not intergrade, but act as separate species toward each other. I have already mentioned a case of this in *Desmognathus,* in a chain containing only three links. It is even possible that there may be more than one meeting of links; the "Formenkreis" may be a "Formenhelix." It is possible to think of the amphibian genus *Caecilia* as a chain of forms containing 16 links; a chain so twisted on itself that as many as five links may overlap. Here the "circle of forms" would be a "spiral of forms with five turns." Certainly five forms of *Caecilia* may occur together without local intergradation. This is true in western Ecuador, for which I give figures, the first column of which is the mean of the total number of body segments, the second the mean of the number of body segments (posterior) with bony scales, the third the mean of the ratios of length to diameter. I know of forms intermediate between each of these five. These forms exist outside of western Ecuador. I have been studying this genus and accumulating data on it for some 14 years. As specimens have come in (324 at

present), gaps between forms have been filling up, but it is ever more clear that several distinct forms inhabit any given area. I can predict with confidence that with double the present material *Caecilia* would be a monotypic genus, with 16 races, as many as five of which might occur together and remain distinct.

Caecilia of western Ecuador

guntheri	118	8	31
dunni	123	38	41
nigricans	168	46	56
pachynema	170	3	64
bassleri	228	20	123

Gradual and regular alterations in characters within a species from one area to another have long been known and have recently been called "clines." These have always been a source of difficulty in systematic work and have always been suspected of being not genetic but environmental. Treated diagrammatically they agree best with diagram 1,a above. Treated taxonomically, it would be better to recognize only the two ends of the series as subspecies and to treat the entire group of intermediates as intergrades and leave them unnamed. Unfortunately in actual practice we seldom meet with a cline without previous taxonomic treatment, and often some one of the middle group has been named; thus it may be necessary to recognize the middle as well as the two ends as subspecies.

Isolating Mechanisms

In cases of geographical subspecies the isolating mechanism is probably sheer distance, plus whatever selectional effect may result from different conditions in different areas.

In ecological subspecies there would be some spatial separation, and a strong selectional effect resulting from the different conditions in different habitats. The best case I know for further investigation is the relationship between *Desmognathus fuscus* and *D. carolinensis*, salamanders of the southern Blue Ridge. This case was mentioned previously. It has the double interest of being an ecological subspecies in *statu nascendi*, and of being two links in a chain of three forms of which the two ends (*D. fuscus* and *D. ochrophaeus* of the Appalachian Plateau) seem completely distinct from each other.

I mention, parenthetically, another interesting case in the same genus and in the same region. This is the "*imitator*" variety of *D. carolinensis*,

which seems to differ from *carolinensis* by two genetic factors, and which occurs only in the Great Smokies within the range of *Plethodon jordani* which it resembles exactly in color. This is the only known case of possible "mimicry" in Amphibia and deserves further investigation.

A case of physical isolating mechanism may be the dwarf *Desmognathus wrighti*, occurring entirely within the range of *D. carolinensis*, and apparently derived from it. This may be an example of specific isolation *ab initio*. If the dwarfing took place at a single step, as is not impossible or unlikely, it is highly improbable that the derivative would be able to mate with the parent form, and would be able to mate only with similar dwarf mutants. Thus a dominant mutant of this sort would almost certainly never reproduce, but a double recessive mutant of this sort would occur with others in the same brood and some possibility of mating would be given.

The recent work of A. P. Blair on mechanisms of hybridization and isolation in toads and in tree-frogs is an excellent example of work along these lines, bringing out indications of ecological, temporal, psychological, and physiological mechanisms. There is a puzzling condition in New Jersey swamp frogs, *Pseudacris*, segregation of individuals into two sets with different calls having been observed at Moorestown. There is seasonal overlap but some seasonal difference between the two voices; there is a slight difference in marking between the individuals corresponding to the two voices; but the striking thing was the marked separation of the breeding individuals in nature, one type of voice not emanating at all from the area of the other and vice versa.

GENERA

The arrangement of the discrete unit groups (species) into higher categories is done on a basis of morphological similarity. The original purpose was to afford a convenient classificatory basis as an aid in identification and in reference. Since 1858 the arrangement has endeavored to convey implications of relationship. These two considerations, convenience and relationship, are, or should be considered to be, the criteria of genera. There have been attempts from time to time to set up more objective criteria, usually morphological. To the herpetologist, Cope's attempts along this line are well known, at least by results. Cope was frank to admit that such criteria, strictly applied, might produce groupings which indicate particular morphological categories, rather than groups of related species. The erratic variability of characters is the great objection to such criteria for genera. The loreal plate

is characteristic of many snake genera, its absence is equally characteristic of many other snake genera (almost so of the family Elapidae), but I cannot see that the type of *Synchalinus coralloides* Cope from Costa Rica is anything but a young specimen of *Pseustes poecilonotus*, which, as an individual variation, lacks this plate.

The smallest supraspecific group possible would include a single species ("Formenkreis") with its races (if any), a monotypic genus. Such groupings may be necessary at times; they emphasize the differences of the group from other groups, but as they do not indicate the relationship of the species with any other species, they tend to defeat the original purpose of binomial nomenclature. No criteria sanction placing races of the same species in different genera.

The second smallest grouping would be a "supraspecies" or "Artenkreis," a series of vicarious forms, some distinct and some intergrading, morphologically and ecologically similar, and manifestly of common origin. It should seem obvious that by the criteria of relationship and of convenience members of the same "Artenkreis" should be members of the same genus. Yet the monotypic genus *Salamandrella* is still current usage for one aberrant member of the Asiatic mainland *Hynobius*, and the monotypic *Tropidoclonion* for an aberrant member of the *Thamnophis elegans* "Artenkreis."

Under the heading of convenience the matter of ease and certainty of generic allocation is important. There are numerous herpetological "species" which would never have been described had not fallacious and difficult generic "characters" been overemphasized. One needs only to mention the grooving on the hind teeth of snakes, or the shape of the tentacular aperture in caecilians.

Also under convenience might be considered the practical value of a group name for forms which are ecologically similar, and which are related or similar morphologically. "Formenkreise" and "Artenkreise" are such groups almost by definition, but many more complex groupings than these exist in nature. Since a marked difference in ecological status usually entails a marked morphological change, the herpetological genus of best current practice, although based on morphology, also expresses the ecology.

Generic names are of value if and when they are useful. Utility means that the species included: (1) are related, (2) are easily and obviously separable from species of other genera, and (3) are similar ecologically. Problems arise when transitional species exist, as they often do. These, to our embarrassment, prevent easy and convenient diagnosis of

supraspecific groups. It might be preferable to refer to such connected groups of species as sections of genera. To name them as genera is to confer on them an aura of distinctness and separateness which they actually lack; to name them as subgenera is to add unnecessarily to the nomenclature and give a temptation to a later and perhaps less well informed student to raise them to generic rank.

CRITERIA OF SUBSPECIES, SPECIES AND GENERA IN ORNITHOLOGY*

By

Ernst Mayr

The American Museum of Natural History, New York, N. Y.

It is an impossible task to give an adequate discussion of the criteria of subspecies, species and genera in the short time of 15 minutes. So let us lose no time in preliminaries but start immediately with the discussion of *species*, which after all is the most important of all taxonomic categories. The species concept has become increasingly confused in recent years by the application of rigid and arbitrary criteria. Some authors determine the status of a natural population with the help of *genetic* criteria; others, through *morphological* criteria, such as lack of intergradation or degree of morphological difference; still others, on the basis of diminishing *fertility*. In determining whether or not such criteria are valid, we must go back to the fundamentals of the species concept. Let us first ask: *What is a species?*

The species concept is almost as old as mankind itself. I once spent several months with a primitive tribe of Papuans in the interior of New Guinea and found that these people had a different vernacular name for nearly every species of bird occurring in their territory. What these natives distinguish as different kinds of birds corresponds exactly to the species of the taxonomist. The same is true for our local birds. The field naturalist recognizes, for example, five kinds of thrushes of the genus *Hylocichla* in the American Northeast the wood-thrush, veery, hermit thrush, gray-cheeked thrush and olive-backed thrush Their ranges overlap broadly; in fact, up to three of these kinds of birds may nest in the same woods, but no intermediates or hybrids are known. Furthermore, these five species differ in their songs, courtship habits, nests, migratory habits and about every other attribute that has been studied carefully. Local populations within any of these five species are *interbreeding*, but each of the five units is completely separated from the others; it is *reproductively isolated*, in the terminology of the modern biologist. And thus we have arrived at two of the basic criteria, in fact, at *the* two basic criteria of species first, interbreeding of the local populations belonging to the species; second, reproductive isolation of those populations which do not belong to the same species.

*A more detailed survey of this field is presented in the author's book, "Systematics and the Origin of Species" (Columbia University Press)

After the museum taxonomist obtained specimens of what the field naturalist recognized as species, he discovered that each of these interbreeding units, called species, showed certain morphological characteristics which permitted the easy identification of dead specimens. These characteristics, the so-called taxonomic characters, gained an ever increasing importance, particularly when it concerned the classification of specimens from countries or systematic groups about which field naturalists were unable to provide pertinent data. Eventually most taxonomists forgot that species were aggregates of living organisms; they forgot that in the museum they were dealing merely with dead samples of the true species of nature, and one of the consequences of this working technique of the taxonomist was that the taxonomic character advanced from a convenient handle in practical work to the level of an absolute criterion. The conclusion, "the level of reproductive isolation—that is, the species level—is associated with certain morphological differences," was reversed to read: "a certain degree of morphological difference proves reproductive isolation—that is, it proves specific difference." This conclusion is one of the most famous fallacies of logic. The ancient Greeks illustrated it in the form of a so-called "vicious syllogism".

(1) Birds have two legs.
(2) Man has two legs.
(3) Therefore, man is a bird.

In our case the fallacious syllogism reads:

(1) Forms that are known to be good species are separated by morphological gaps.
(2) Forms A and B, whose specific status is in doubt, are separated by a morphological gap.
(3) Therefore, A and B are good species.

This, of course, is true only in some cases, not in others. The absurdity of the strictly morphological criterion would become perfectly evident if we were to apply it to males and females in sexually dimorphic species, or if we were to recognize larval or immature stages as separate species or give specific rank to the various forms of a polymorphic species. As a matter of fact many "species" of the taxonomic literature have turned out to be nothing but such intraspecific variants. However, as soon as it was proven that they interbreed with other variants, they were deprived of their specific rank, even though they remained separated from each other by clear-cut, unbridged gaps in morphological characters. To repeat once more, in a given locality the criterion of interbreeding

versus reproductive isolation will permit us to determine, in nearly all cases, which individuals belong to one species and which to another. Morphological differences are of practical convenience, but not a primary criterion.

What is true for contiguous individuals—Poulton calls them *sympatric* individuals or species—is equally true for those that do not occur in the same geographical districts, the so-called *allopatric* forms. No degree of morphological distinctness is in itself proof of specific distinctness. The probability of reproductive isolation is the primary criterion. To emphasize this point is important, because the degree of reproductive isolation of a geographically isolated form is not necessarily correlated with the degree of morphological distinctness, nor is the degree of morphological distinctness necessarily a good index of genetic distinctness. A single gene difference may produce a phenotypic difference which might cause some taxonomists to call the population carrying it a different species. This consideration again leads us to the inevitable conclusion that a species concept, based primarily on the criterion of morphological distinctness or of a gap in morphological characters, is invalid.

Please do not misunderstand me. I am not demolishing the morphological species concept merely because I enjoy being an iconoclast. No, I attack it because I consider it contrary to the fundamental concept of species. But, you will ask me, how should we treat geographically isolated populations? Should we follow Kleinschmidt and call *all* of them subspecies? Such a procedure would have the advantage of consistency, but actually it is just as unscientific as calling all morphologically separated populations species, because there is little doubt that many of these isolated populations have already reached the species level. But how shall we decide whether or not these allopatric forms are species?

A complete analysis, including experiments on mating preference and a cytological examination of possible hybrids, is required in order to reach a satisfactory decision. To make such a painstaking analysis is out of the question in practically all the cases which interest the taxonomist. The taxonomist is forced, in most cases, to determine the status of a geographically isolated population by indirect methods. Reproductive isolation is correlated with a certain degree of morphological difference, which is rather typical for any given genus. It is very small in the flycatcher genus *Empidonax* or in the mosquito genus *Anopheles*, but extraordinarily large among the birds of paradise. If we want to determine the taxonomic status of a geographically isolated population, we must first study all the species of the genus to which this population belongs,

and the species of related genera. In this manner we can work out a scale of differences between unquestionably valid species and between unquestionable subspecies. The status of the doubtful population will have to be decided by inference.

In conclusion I might say that the species concept, as just set forth by me, is by no means endorsed by all ornithologists. Many of them consider, for example, the yellow-shafted and the red-shafted flicker, or the Oregon and the eastern junco as separate species, even though individuals of the respective forms seem to interbreed indiscriminately wherever they come into geographical contact. A proponent of a biological species concept has no choice but to consider such completely interbreeding forms as conspecific. So much for the species.

The *subspecies* is composed of a group of local populations and can be distinguished from other such groups by one or several taxonomic characters. There are three principal difficulties involved in the recognition of subspecies.

(1) *How can it be determined whether an isolated population is a subspecies or a species?* This question was just discussed in regard to the species and no further comments are needed.

(2) *Is the geographical race the only kind of subspecies?* Invertebrate and plant taxonomists recognize ecological races. Recent research indicates that some of these so-called "races" are purely phenotypical, while others are microgeographical races No such ecological races of birds have been described. It might be emphasized, however, that every geographical race owes a greater or smaller proportion of its characteristics to the selective influences of the particular local environment. Every geographical race is, thus, to a greater or lesser degree also an ecological race. In some cases, as for example on black lava flows, geographical races develop which owe their most conspicuous (and often only) taxonomic character to the selective qualities of this particular habitat. I see little advantage in calling such forms ecological races rather than geographical races, because each of these races has many characteristics which are not the result of local selective factors. It will be the task of the student of plants and invertebrates to determine whether or not it is possible to recognize the ecological race as something quite distinct from the geographical race. One point should not be overlooked in such an analysis. Whenever two ecologically differentiated races overlap in the same locality, a careful analysis of the situation usually shows that this is a secondary condition and that the original ecological difference developed during a preceding geographical isolation;

that is, the race was a geographical race before it became an ecological race.

(3) The third difficulty in recognition of the subspecies can be expressed in the question: *How can subspecies be delimited from each other in continuous populations?* Subspecies borders are generally drawn where there is a distributional gap, or a change of environment, or a significant change in the taxonomic characters of the continuous populations. A detailed analysis of the populations near subspecies borders shows in many cases that the change from one subspecies into the next is so gradual that the placing of these borders is left to an arbitrary decision. In other cases equally arbitrary decisions must be made concerning the degree of difference between two groups of populations which is to be considered sufficient for separation of subspecies. There are splitters and lumpers. In general it is stated, in the ornithological literature, that 75 per cent of the individuals of one race must be clearly separable from all the individuals of the other race. This is a very unsatisfactory method of handling the problem, since the status of many subspecies changes with the increase in number of collected specimens. A way should be found to express the necessary degree of distinctness in more absolute terms, such as standard deviations. Beginnings are being made along these lines, but they have not yet found their way into the taxonomy of birds.

And now for the *genus!* Recent historic research has proven conclusively that the genus of Linnaeus and his forerunners goes back to an equivalent concept of folk-lore. There was a concept "oak," before there was a genus *Quercus*, and a concept "finch" or "thrush," before there were any such genera as *Fringilla* and *Turdus*. Up to this point the genus presents a parallel case to the species. However, whereas the species of Linnaeus, in the great majority of the cases, still corresponds to the species of today, the genus of Linnaeus only rarely does so. In most cases it is now equivalent to a subfamily, a family, or even an order. There is little argument among contemporary ornithologists concerning the delimitation of species, but the divergence of opinion in regard to the genus is tremendous. The genus of such splitters as Mathews, Roberts and Oberholser has little in common with the genus of lumpers as represented by Stresemann, Peters and Mayr, to mention some of the contemporary ornithologists. Let us illustrate the difference between splitters and lumpers by examining a part of a phylogenetic tree.

Let us assume we have two phylogenetic branches, A and B. Branch A again will break up into four or five twigs and twiglets. Both lumper and splitter agree that all the twigs on branch A are derived from a com-

mon ancestor; they also agree that there is a definite gap between A and B, and that all the species of A have certain characters that distinguish them from all the species on B. However, the splitter will emphasize the reality of the little gaps between the twigs on branch A, while the lumper will contend that the recognition in nomenclature of all minor gaps will obscure rather than clarify the true relationships and the basic pattern of classification.

We now must ask, what are generic criteria?

It is perhaps not always realized or at least not kept in mind that subspecies, species and genus differ from all other systematic categories by being the only ones that are mentioned in the scientific name of an individual animal. The significance of this nomenclature must therefore be examined if we want to consider the status of the categories which they represent. We can eliminate the subspecies from our discussion, since what is true for the species is also true for the subspecies, and concentrate on genus and species. There is a fundamental difference between these two categories. The species is an individual unit; the species name therefore emphasizes distinctness.

The genus, on the other hand, is a collective unit and the joint application of the same generic name to a number of species indicates their similarity or relationship. The functions of the two components of the scientific name as proposed by Linnaeus are therefore diametrically opposite. The species name signifies singularity and distinctness, the generic name implies the existence of a group of similar or related units. This difference in the functions of species and genus names is completely ignored by many recent taxonomists, particularly the so-called generic splitters. It is their aim to express difference not only in the specific, but also in the generic name. This tendency, if carried to its logical extreme, leads to uninomialism, and some of the leading generic splitters have openly or in a veiled form endorsed this principle of nomenclature. To me it seems to indicate a complete misunderstanding of the principle of binomial nomenclature, if somebody uses the generic name primarily to express difference. This is the function of the species name. The generic name was introduced by Linnaeus into nomenclature in order to relieve the memory and this should remain its principal function. Wherever the genus becomes too small, it loses its usefulness.

The splitting fever has played havoc in bird taxonomy. We have now more than 10,000 generic names for an estimated 8500 species of birds. Even conservative authors admit 2600 genera, which amounts to only 3.27 species per genus of birds. An average of five species per genus would bring the number of bird genera down to about 1700, which would

be within the capacity of the memory of a single human individual. The trend toward larger genera has been unmistakable in ornithology and it has been accelerating in recent years. I do not know of a single younger author who could be classed as a generic splitter.

The genus should contain groups of similar species—that is, species which we consider related—but so far as I know nobody has ever found an *objective* criterion of the genus. Personally, I like best a genus definition which is based on honest admission of the subjective nature of this unit. It may not be possible to go much beyond the definition of the entomologist Thorpe who said: "The genus, to be a convenient category in taxonomy, must in general be neither too large nor too small." A more comprehensive definition would be: "A genus is a systematic unit including one species or a group of species, presumably of common phylogenetic origin, which is separated from other similar units by a decided gap. It is demanded for practical reasons that the size of the gap be in inverse relation to the size of the unit." This latter qualification will prevent the recognition of many monotypic genera.

We can summarize this discussion as follows: The taxonomic category of the genus is based on the fact that species are not evenly distinct from one another, but are arranged in smaller or larger groups, separated by smaller or larger gaps. Recognition of the genus is, therefore, based on a natural phenomenon. How many of such groups are to be included in one genus and how the genus should be delimited from other genera are matters of convenience left to the judgment of the individual systematist. Taxonomic characters that prove generic distinctness do not exist. Taxonomic literature could have been spared many unnecessary generic names if the taxonomists had kept in mind Linnaeus' warning: "The characters do not make the genus, but the genus gives the characters."

One more word in conclusion—all taxonomic categories are collective units. The subspecies, the species and the genus all consist of groups of unequal components: the subspecies of local populations, the species of local populations and subspecies, the genus of more or less distinct species. Those authors who try to obtain homogeneous units by splitting them down to their last elements are not only bound to fail in their endeavor, but they also obscure the basic relationships.

CRITERIA FOR VERTEBRATE SUBSPECIES, SPECIES AND GENERA: THE MAMMALS

By

E. Raymond Hall

University of California, Berkeley, California

Mr. Chairman, members of the American Society of Ichthyologists and Herpetologists, members of the American Society of Mammalogists, and guests: We had expected as a speaker at this time one of the senior mammalogists who now is unable to attend. I am glad to appear as a substitute because the subject under discussion is one in which I am especially interested. In these extemporaneous remarks I propose: (1) to indicate some steps which I think useful to take in classifying mammalian specimens as to subspecies; (2) to express my personal views as to criteria for subspecies, species, and genera of mammals; (3) to illustrate how some of these criteria for subspecies and species may be applied to closely related insular kinds of mammals; and (4) to suggest a way in which subspecies may disappear without becoming extinct.

When I undertake to classify mammalian specimens as to subspecies or species, or when I present a series to a beginning student for classification, I like to observe the following steps: (a) select for initial, intensive study a large series, 30 or more individuals, from one restricted locality; (b) segregate these by sex; (c) arrange specimens of each sex from oldest to youngest; (d) divide these into age-groups and within a given group, of one sex, from one locality, of what is judged to be one species, measure the amount of so-called individual variation; (e) with this measurement as a "yardstick," compare individuals, and if possible series, comparable as to sex and age (and seasons where characteristics of the pelage are involved) from this and other localities. The differences found are usually properly designated as geographic variations and form the basis for recognition of subspecies, which in turn comprise one of the tools used by some students of geographic variation.

As to criteria for the recognition of genera, species and subspecies of mammals, it seems to me that if crossbreeding occurs freely in nature where the geographic ranges of two kinds of mammals meet, the two kinds should be treated as subspecies of one species. If at this and all other places where the ranges of the two kinds meet or overlap, no crossbreeding occurs, then the two kinds are to be regarded as two distinct, full species. The concept of a species, therefore, is relatively clear-cut

and precise; the species is a definite entity. Furthermore, if a zoologist knows the morphological characteristics diagnostic of the species, he has no difficulty in identifying a particular individual as of one species or another. In identification of subspecies, difficulty is frequently encountered, especially with individuals which originate in an area of intergradation.

The category next higher than the species, namely, the genus, is less definite and more subjective as regards its limits than is the species. As the species is the definite, clear-cut starting point for defining subspecies, the species is likewise the starting point for consideration of genera. Degree of difference is the criterion for a genus. The genus lies about midway between the species and the family. Because the limits of the family, like those of the genus, are subjective, it follows that the criterion for recognition of genera, although precise enough at the lower point of beginning, the species, is elastic at the upper end—namely, at the level of the family.

In summary, the criterion for subspecies is intergradation, that for species is lack of intergradation, and that for genera is degree of difference. These ideas agree in general with the ideas expressed by the previous speakers.

One of the situations in which it is difficult, or impossible, to apply these criteria to conditions actually existing in nature is comprised in some insular populations. Frequently the populations on two islands near each other differ enough to warrant subspecific or possibly specific distinction. A means of deciding on specific versus subspecific status for these populations is to find on the adjacent mainland a continuously distributed, related kind of mammal which there breaks up into subspecies. Ascertain the degree of difference between each pair of mainland subspecies which intergrade directly. If the maximum degree of difference between the insular kinds is greater than the difference between the two subspecies on the mainland, which intergrade directly, and greater than that between either insular kind and the related population on the nearby mainland, the two insular kinds may properly be treated as full species. If the maximum degree of difference between the insular kinds is no greater than, or less than, the difference found on the mainland between pairs of subspecies which intergrade directly, the insular kinds may properly be treated as subspecies of one species. In fine, the criterion is degree of difference with the limitation of geographic adjacency, rather than intergradation or lack of it.

Now to my fourth point, namely the suggestion that many subspecies disappear without becoming extinct. Permit me first to observe that

although species and subspecies seem to have the same kinds of distinguishing characters, which appear to be inherited by means of essentially the same kinds of mechanisms in the germ plasm, there are two noteworthy differences between species and subspecies. One already implied is that, in a species which is continually distributed over a given area, its characters at the boundaries of its range are sharp, definite, and precise. Some of its characters comprised in size, shape and color, at any one place are either those of one species or instead unequivocally those of some other, whereas the characters of a subspecies, particularly at or near the place where two subspecies meet, more often than not are various combinations of those of the two subspecies and in many individual characters there is blending.

Second, through a given epoch of geological time while a species is in existence, one or more of its subspecies may disappear and one or several new subspecies may be formed. Subspecies, therefore, on the average are shorter-lived than species.

Now the disappearance of subspecies is to be expected on *a priori* grounds if we suppose that new subspecies are formed in every geological epoch. There is reason to believe that in the Pleistocene, the epoch of time immediately preceding the Recent, there were even more species of mammals than there are now. In each of several successively corresponding periods of Tertiary time before the Pleistocene, probably there were as many species as now. Probably too, these species then were about as productive of subspecies as species are now. Had even half of these subspecies persisted, either as subspecies unchanged or in considerable part by becoming full species, there would now be an array of species and subspecies many times as numerous as actually does exist. It is obvious therefore that many disappeared.

In accounting for this adjustment of numbers of kinds of mammals, I have spoken of the disappearance of subspecies rather than of their extinction because I can imagine how a species, say, the pocket gopher *Thomomys townsendii*, in the middle Pleistocene with three subspecies (geographic races) could have come down to the present by means of each of the three subspecies having gradually changed its characters into those of one of the three subspecies existing today in the area of northern Nevada that I have in mind. In this way, disappearance of subspecies living in the Pleistocene has been accomplished, without their having become extinct in the sense that the subspecies left no living descendants. Of course this has to be true for some of the subspecies of each successively preceding epoch if any animals at all persist, but what I wish to emphasize is the strong probability that many, perhaps more than 50 per cent,

disappeared thus without actually becoming extinct, when, for example, two successive stages of the Pleistocene, south of the ice sheet, are considered. In this regard it is pertinent to recall that each of three Pleistocene kinds of pocket gophers, *Thomomys* (probably species *talpoides*) *gidleyi*, *Thomomys* (probably species *townsendii*) *vetus*, and *Thomomys* (probably species *bottae*) *scudderi*, from a short distance over the northern boundary of Nevada, differs from living representatives corresponding to it (several subspecies of one species) in greater width labially of the individual cheek teeth of the lower jaw. Significant for the thesis being defended is the point that each and all of these *Thomomys* in the Pleistocene differed, at least as regards the shape of the teeth, in the same way from the three living species which I feel confident are their descendants.

Let us suppose that three hypothetical subspecies of *Thomomys townsendii* in middle Pleistocene time each gradually changed into three different subspecies inhabiting about the same areas in upper Pleistocene time, and that these in turn were the ancestors of the three subspecies living in those same general areas today. A total of nine kinds is thus accounted for. At any one time there was geographical intergradation, which has reference to horizontal direction. Also there was intergradation up through time, which has reference for present purposes to a vertical direction. If I had before me all the material necessary to substantiate this or a similar case, I would be inclined to recognize nine subspecies of one species. This hypothetical case emphasizes the importance of intergradation, the criterion for subspecies.

In review: I have mentioned some preliminary steps useful for a person to take when he aims to analyze variation in mammals and to establish species and subspecies thereon; intergradation is the criterion for subspecies and degree of difference is the criterion for genera; degree of difference with the limitation of geographic adjacency may be used as the criterion for insular populations (the classification of which is doubtful as between subspecies and species); and, finally, I have sought to stress the importance of intergradation as a criterion for subspecies by showing how subspecies may disappear without becoming extinct.

CRITERIA FOR GENERA, SPECIES, AND SUBSPECIES IN ZOOLOGY AND PALEOZOOLOGY

By

George Gaylord Simpson

The American Museum of Natural History, New York, N. Y.

INTRODUCTION

This paper does not present the consensus of vertebrate paleontologists, nor is it confined to vertebrate paleontology. The criteria used by various students, or by each of them at various times, for recognizing genera, species, and subspecies among fossil vertebrates are so different that it would be a gigantic task to summarize them all and an impossible task to find among them procedures in universal use. It is certainly more practicable and it may be more useful to approach the problem more broadly and at the same time more individually. This study, then, presents an individual opinion as to the general principles that do or that should underlie the selection and use of criteria for these taxonomic categories.

The viewpoint is paleozoological throughout and the differences between paleozoological and neozoological problems and criteria are discussed, but these two branches of zoology have much in common and a search for general principles cannot be confined to one branch. Although there is thus some invasion of the fields of other contributors to this symposium, it may be hoped that even here the difference of viewpoint will have some value and will prevent mere repetition.

DEFINITIONS

The Species

The title of this symposium was deliberately chosen to avoid the direct question "What is a species?"—a subject of such long, frequently futile, and sometimes acrimonious debate. The theme of the present discussion is not what genera, species, and subspecies are in a theoretical sense, but how specialists in the different branches of vertebrate zoology should set about recognizing and delimiting such units in the practice of classification. The distinction is not merely verbal. One of the points that I want to emphasize is that the species in nature is something different

from the species in classification. It is, however, impossible to consider criteria for recognition of a taxonomic group without having a reasonably clear idea of the nature of such a group. Assuming, for the present, agreement with the belief that the species is the fundamental unit of classification, its nature may first be discussed and those of subspecies and genera may be related to the concept of species. Categorical statement is necessary for brevity and does not indicate lack of awareness of radically different opinions or unwillingness to grant validity to other points of view.

A species in nature is a group of organisms. It is not a process, as some geneticists say; or an infinite mathematical abstraction, as some statisticians maintain; or a collection of individual specimens, as no one is likely to say but as many, perhaps most, working taxonomists seem unconsciously to assume. The group arises by dynamic genetic processes, it can be described and interpreted by statistical methods, and it is composed of individuals, but it is the group itself that is a species.

There are not literally an infinite, but certainly an enormous number of possible and real groups of organisms differing both in content and in kind. In modern taxonomy it is a basic concept that the species in nature is a genetic group. The kind of genetic group that should be called a species grades into kinds that are given other names and this gradation, sometimes denied, is the most fruitful source of misunderstanding and disagreement. Nevertheless one idea underlies most of the definitions given by evolutionary taxonomists and is clearly involved in recent discussions by such an able zoologist as Mayr (1940, p. 256) and by such an able geneticist as Dobzhansky (1941, p. 373). The idea is expressed in various ways and with different qualifications and exceptions made necessary by special situations, but it is fundamentally this: *a genetic species is a group of organisms so constituted and so situated in nature that a hereditary character of any one of these organisms may be (possibly, but not necessarily) transmitted to a descendant of any other.** To the paleozoologist there is a large, important field in which definitions involving this idea are inapplicable: vertical species, dynamic temporal sequences. Discussion of this point is deferred to a later page, and for the present consideration is limited to the field in which this genetic concept is appropriate.

The neozoologist could conceivably apply this concept directly to the observation of nature and use it as his criterion for species in classification. In reality he does no such thing. What he does instead, I leave

*Obviously a primary qualification is that definitions implying this can apply only to sexually reproducing organisms. Since all vertebrates normally reproduce sexually, discussion of asexual or parthenogenetic species is not pertinent in this symposium.

to the other contributors to this symposium, except as his practice is related to that of paleozoology. For paleozoologists the direct use of such a criterion is not conceivable, even aside from the difficulty pointed out in the last paragraph. The neozoologist, by custom and for practical reasons, and the paleozoologist, from necessity, both define their species by morphology and not by the transmission of heredity or breeding habits and potentialities. Again subject to exception and modification for special cases, and again inapplicable to successive stages in vertical sequences, a species may be defined morphologically more or less as follows: *a morphological species is a group of individuals that resemble each other in most of their visible characters, sex for sex and variety for variety, and such that adjacent local populations within the group differ only in variable characters that intergrade marginally.*

This morphological definition is merely a description of the usual result of the situation involved in the genetic definition. Therefore, morphological species tend to correspond closely to genetic species, although it cannot be expected that the correspondence will be exact and universal. This treatment of present theory really reverses the historical sequence. The objective effect of the genetic situation was observed long before there was any clear idea of that situation or of phylogenetic processes in general. Species were defined morphologically before the concept of a phylogenetic species was achieved, but just because the species are real groups and because the phylogenetic situation does have definite morphological effects, it turns out that essentially the same groups are called species under both definitions.

Now that it is admitted that the phylogenetic unit, the species of evolutionary theory, is best defined by breeding or genetic structure, the morphologically defined species is not to be considered the species proper, in the strictest sense, but it is what most classifiers agree to call a species. This does not at all mean that the morphological species is a subjective or artificial concept. It is a real group that nearly corresponds with and is taken as a sufficient approximation of the genetic species.

Even the morphological definition does not bring us down to what the taxonomist really observes or to the species that he actually defines. What is observed and described is a series of individual specimens. These specimens do not constitute the species, and their description and differentiation from other series certainly do not constitute the description or diagnosis of a species. That this is what nine authors out of ten call a definition of a species does not alter the fact that they are mistaken or that they deal with species only by implication and not in their

literal expressions. The series of specimens in hand is a sample drawn from a natural population (here assumed to be a species) but never completely representative of that population. If a species is defined, the process is to infer from the sample the characters and limits of the morphological species from which the sample was drawn. This inference is the species that really is used in taxonomy. The taxonomic species is a subjective concept and it cannot be exactly equivalent to the morphological species (which is an objective group), but the taxonomic species approximates the morphological species more or less closely according to the adequacy of the sample and the skill of the taxonomist. Thus we need a third definition of a species: *a taxonomic species is an inference as to the most probable characters and limits of the morphological species from which a given series of specimens has been drawn.*

What is really done in classifying organisms is to base, on a series of specimens, a taxonomic species which is a subjective estimate of a morphological species, which in turn is a group of organisms so defined as to approximate a genetic species. Practical classifiers will probably object to the complexity of this statement, but the complexity is only revealed, not created, by the analysis. This is the process and it is complex.

The Subspecies

It is true of all other categories of classification, as of species, that three distinct entities are involved, one taxonomic, one morphological, and one phylogenetic. There are three main sorts of categories: the species, the subspecies, and all higher categories, from "species groups" or subgenera up to kingdoms.

Genetically a subspecies may be approximately defined as a group of organisms throughout which active interbreeding occurs regularly, or so that the average hereditary repertory is approximately the same in the various local subgroups. This is by no means as clear-cut a thing as a species, even granting that the latter also shows considerable intergradation. Even the smallest local groups, below any level reasonably called subspecies, almost always show some differences in average genetic composition and adjacent "good" subspecies do usually interbreed normally where they are in contact. There are species in which fairly homogeneous* local groups have well-marked genetic differences from their neighbors and intergrade with the latter only along narrow zones

*Homogeneity does not mean that all individuals are alike; on the contrary they may be markedly different within a homogeneous population. It means that any two samples drawn from the population at random will not differ significantly in composition, or in representation of diverse variants. An analogy: a mixture of salt and pepper is homogeneous if the proportion of salt to pepper is about the same throughout. there is no implication in the word homogeneous (as here used) that it must be all salt or all pepper.

of contact. Here there is little doubt as to the reality and extent of the subspecies (provided the genetic facts are known). On the other hand, there are species in which only the most restricted groups—point rather than area groups—are at all homogeneous and the intergradation is essentially continuous from one end of the species to the other, although the ends may be quite different. In such groups, the clines of Julian Huxley (1938), there is no natural number of subspecies and no way to delimit them that is not arbitrary. The situation has an interesting analogue in continuous vertical sequences, to be discussed later. The practice in such cases is to make arbitrary divisions such that the differences between their median characters are approximately of the order of those between subspecies with limits observable in nature. An alternative, perhaps more logical but as yet little used in practical classification, would be to define the terminal points and to designate intermediate conditions by their distance from the ends. Finally there are species in which the whole population is fairly homogeneous. In such circumstances the species is commonly said to lack subspecies. Logically, it might be preferable to say that the species then includes only one subspecies, an example of monotypy, which also has special importance for the paleontologist and will be discussed later.

The relationship of taxonomic and morphological to genetic subspecies is closely analogous to that of taxonomic and morphological to genetic species and is sufficiently obvious on that basis. Given the difference in genetic definition, the important distinction between species and subspecies in practical work lies chiefly in the differences in methods of inference from specimen-series to taxonomic species and subspecies.

The Genus

The basic element in phylogenetic subspecies and species is a form of genetic continuity, actual or potential; that is, hereditary characters are being or can be transferred from one part of the population to another. By definition, such transfer cannot normally occur between different species,* and above the rank of species there may be genetic discontinuity within the group defined as a taxonomic unit. These higher categories are all essentially alike in this respect. They differ from each other in scope and, for the paleontologist, to some extent in their balance between horizontal (geographic) and vertical (time) dimensions. Thus in theory there is an absolute distinction between species and genus, and the dis-

*It is not necessary here to take up the old question of fertile interspecific and intergeneric hybrids. The isolating mechanism does not have to be reproductive but can be geographic, mechanical, psychological, etc. *Bison* and *Bos* can produce fertile hybrids, but they did not, in fact, interbreed in the undisturbed natural conditions under which the genera arose.

tinction can usually be made in practice, but there is no theoretical qualitative difference between genus and family. The two intergrade and it is a matter of custom and taste where the line is drawn.

According to universal agreement, a genus is one of the lower supra-specific categories. It includes either a cluster of species of not long antecedent common origin, or a single species that differs as much from other known species as if it were more or less central in a cluster of which the other members are missing. This is a vague definition and provides no rule of thumb, but no better can be expected. Unlike a species, the genus is a member of a continuously intergrading hierarchy of categories and therefore cannot be defined exactly. This does not mean that a genus is not "real" or "natural," but that the distinction of one member of the hierarchy as a genus requires art as well as science.

Some students insist that the genus be the lowest category employed above a species. The *reductio ad absurdum* of this principle is to place every species in a separate genus, a tendency that is operative in the work of a number of specialists. To name only one of many examples, the successive revisions of Severtzow (1858), Pocock (1917), J. A. Allen (1919) and others, resulted in recognizing at least 23 genera of living felines and placing every well-established species in a different genus.* Obviously, such classification makes the genus useless as a category in taxonomy. On the other hand, some students tend to use the genus for the largest group of species that can be demonstrated, with reasonable certainty, to have a common and exclusive ancestry. Placing all living felines in the single genus *Felis* is an example of this. The arguments for lumping and splitting will probably be covered adequately by the neo-zoologists on this program and it suffices for me to suggest that both extremes should be avoided, but that the over-inclusive use of the genus is less harmful and more acceptable than the tendency to equate it with the species.

Well-balanced use of the genus does leave a practical need within some polytypic genera for a lower collective (supra-specific) category. The need is well filled by the use of subgenera. There has been an unfortunate recent tendency to neglect the subgeneric category, but there are signs that it is being revived either as such or in the form of "species-groups," "supra-species" and the like.

---

*These authorities did, it is true, list more than one species in some of their "genera," but most of the additional species were either poorly known to them or must be judged by a dispassionate reviser to be subspecies by the best recent standards.

PROBLEMS OF INFERENCE

The erection of a taxonomic subspecies, species, or genus by inferring the nature and limits of corresponding morphological groups from a given series of specimens is essentially a statistical problem. Many zoologists —until recently it would have been fair to say "most zoologists"—are strongly resistant to this statement and maintain that they neither need nor use statistics in their work. However, the need is not open to any question and most zoologists whose taxonomic work is sound are using statistical methods whether they realize it or not. The misunderstanding is not primarily the fault of the zoologists, because it has been fostered by the statisticians. In a narrow sense, to which many statisticians adhere, statistics comprises a certain set of mathematical operations carried out with numerical data. Many of these operations are useful in zoology, but good taxonomic work can be and is being done without them, or with only the most elementary of them. In a broader sense statistics is the science of: (a) estimating the characteristics of populations from samples; and (b) describing groups, as such, rather than individuals taken singly. Since these two things are precisely what a systematist does when he sets up a species (or other taxonomic group) on the basis of a series of specimens, it follows that he is using statistics in a broad sense. Of course his statistics may be good or bad, rational or intuitive.

Although the species is the basic and, despite disagreement, the most easily defined genetic unit, the subspecies, at its best, is the simplest taxonomic unit because it is the one in which there is or may be a direct, simple statistical relationship between sample and population. Given a sample such as to make the assumption permissible, it is assumed that it is a random representation of a population homogeneous in the sense previously defined. It is then possible to estimate the probable characteristics and limits of variation in that hypothetical homogeneous population. In the simplest case, that estimate is a taxonomic subspecies. The usual methods of estimation are more or less familiar to all working zoologists and space need not be taken to discuss them here. They vary from the loose and often mistaken use of intuition, through the usual and pragmatically valid empiricism of experience derived from handling many such samples, to the relatively exact and reliable use of statistical methods in the narrower mathematical sense.

Since the results of such inference are in any case probabilities and not hard and fast limits, properly conservative interpretations of this kind automatically allow for a reasonable amount of heterogeneity in the morphological subspecies of which the taxonomic subspecies is a mental

Present accumulations of data suggest that so-called subspecies so heterogeneous as to make this image importantly false are in most cases not morphological subspecies in the strict sense by clines, or segments of clines. In this case the taxonomic units derived by inference on the hypothesis of essential homogeneity do not correspond with subspecies as areas on the plane of distribution but correspond with points on a line or on a plane. The taxonomic treatment of this rather common situation is still in its infancy, but it is quite clear that these taxonomic points can be used to approximate and define morphological lines and planes just as well as the taxonomic areal unit, called a subspecies, approximates morphological subspecies. Such approximation of clines must be a secondary inference based on two or more primary or point inferences.

If a species includes only one subspecies, or if only one of its subspecies is known, the description of that subspecies describes the whole species. It might then be proper to say that the species does not need to be and indeed cannot be defined in terms of itself and on any objective basis. The only way in which such definition can be made different from that of the subspecies is by attempting to allow for the subsequent discovery *of other subspecies, an attempt that may make us look very wise or very foolish if the discovery is made, but that is unlikely to have much value.* Such monotypic species can indeed be diagnosed. I wish here to establish a distinction between *definition*, a description of the characters and limits of variation of a single group (of any sort or scope), and *diagnosis*, a statement of the differences between adjacent groups. Definition may be said to tell what a thing is and diagnosis to establish what it is not. Etymologically "definition" means to set limits and "diagnosis" means to know apart or know [the differences] between. Obviously diagnosis plays a large role in zoological taxonomy, generally the major role in routine procedure of sorting specimens. Yet definition is more basic and it seems best to concentrate primarily on this aspect. No two organisms are precisely alike and the important point is to determine how much alike they must be to belong to the same group. This is definition, and it is impossible to make a proper diagnosis between groups without first knowing by definition what belongs within each group. It is beside the point that convenience often dictates the publication and subsequent use of a diagnosis rather than of the definitions on which it is based.

The species is the fundamental *genetic* unit, but polytypic species are already secondary or multiple *taxonomic* units. From the point of view of erection of taxonomic species by inference as to morphological species, if more than one subspecies is present the methods of inference do not

differ from those involved in taxonomic genera. This underlies the rather common feeling among zoologists that the difference between species and genera, like that between genera and subfamilies, is simply a matter of scale and convenience, whereas to the geneticist there is a profound qualitative difference. Granting that phylogenetic classification should attempt to approximate the genetic ideal, the problem of the zoologist is to determine whether the different results of like methods of inference are more nearly harmonious with genetic species or genera.

The polytypic species and genera of taxonomy cannot, like the subspecies, be set up by simple statistical inference from a sample to a hypothetical homogeneous population. Such first-order inferences must be made for each of a number of different populations and then a second-order inference must be made from these data to define a larger and admittedly heterogeneous group. It is customary to set up a combined definition and diagnosis of the heterogeneous group by listing characters believed to be common to all its members and not to be common, either alone or in combination, to members of morphologically contiguous groups. This usually suffices for the rapid labeling of specimens, a *legitimate goal but a very limited one. It certainly does not suffice for any real understanding of the nature and significance of the group.* On higher taxonomic levels, especially as these are revealed to the paleontologist, such definition-diagnosis by common and exclusive characters is more clearly revealed as inadequate and sometimes is quite impossible. For instance it is instructive to try to list characters common to all Equidae and excluding all members of any other group This may be possible, but I have been unable to make or to find such a list, although the Equidae certainly form a very real and valid taxonomic and genetic group It is important to list characters in common, but it is equally important to define polytypic groups by the nature and sequence or arrangement (in space or in time) of the differences between the included lesser groups.*

The most practical criterion for judging whether a polytypic morphological group, in which there is no qualitative distinction between genera and species, corresponds more nearly with a generic or specific genetic group in which there is such a distinction, is by intergradation. If adjacent contemporaneous subgroups within a larger morphological group

*A point that may here be mentioned parenthetically, because it seems to be misunderstood in some otherwise excellent recent work, is that a polytypic group does not usually have a mean or average condition in the same sense as a monotypic group The mean in a single population is a point of central tendency around which variations tend to cluster The same figure derived from several different populations has no such significance and may, indeed, be a value that does not occur, or tend to occur, in any of the included groups In a polytypic group a more or less valid analogue of the mean is the "condition in common" and an approximate analogue of variations from the mean is the range of differing means in the different subgroups

differ in their means but intergrade in all known characters, it is a reasonable inference that the subgroups are subspecies and the group a species. If the subgroups do not intergrade in one well-marked character, or preferably in several characters, it is proper to infer that the subgroups are species and the group a genus. Allowance must be made for adequacy of sampling: a gap may represent a missing subgroup that would overlap both of two that do not themselves intergrade. It is also to be remembered that such intergradation of the populations may not be visible in small samples. It is not to be judged on the basis of the sample in hand but on the basis of the taxonomic concept derived from that sample. On the other hand, if only a few characters are available for study, the differentiating, non-intergrading characters of separate species may be missed.

This criterion does not assure full agreement between taxonomic and genetic groups. There are known certainly valid genetically distinct species—that is, populations that do not or cannot interbreed—that intergrade completely in morphology. These are, nevertheless, uncommon and it is not a serious error to reduce two such species, certainly very closely related, to subspecific status or even to fuse them completely in the taxonomic system. Moreover, it seems probable that such a situation can only be temporary. I do not know any example of the converse situation—that is, of two populations definitely belonging to one genetic species but not intergrading in a morphological character, either directly or through intervening groups.* Probably cases can occur, but they must be rare.

In general the use of this criterion will rarely, if ever, result in confusing subspecies or mere variations within a subspecies for species or species for genera, if valid methods of inference are used, but it may result in failure to distinguish valid genera and species or in giving them less than their genetic rank. When it fails, the criterion is on the side of caution, which seems desirable. Of course the number of such failures, which are not failures to recognize good morphological groups but to make these equivalent to genetic groups, will be in inverse ratio to the number of variable characters observed and to the closeness of inference permitted by the samples.

CATEGORICAL RANK OF CHARACTERS

It would be incalculably valuable to taxonomists if characters or, more strictly, the differences between them could be assigned fixed categorical

*There are of course striking unit *characters* that cannot have any intermediate condition and so cannot intergrade in this sense, but the *populations* exhibiting them can still intergrade through a population showing both or all conditions. Such a character is the dextral and sinistral coiling of shells and Crampton's classic study (1932) shows that intermediate populations do occur.

value; for instance, if a difference between individuals of 20 per cent in size could be taken as prima-facie evidence that they belong to different species, or a difference in tooth formula that they belong to different genera. Many zoologists have believed that they could assign such values and have proceeded on this belief. The search is reminiscent of that of the alchemists for the philosopher's stone. It has been an attempt to find an easy way to do something that can, indeed, be done, but not by easy methods.

In the first place, the belief that a morphological distinction of given degree or kind constitutes in itself a fixed rank of diagnostic distinction in taxonomy is inconsistent with modern taxonomic principles because it approaches the problem on inadmissible premises. Morphological differences are used to describe and distinguish taxonomic groups, but it is the groups of organisms that are being classified, not the morphological characters themselves. That the distinction is important and not particularly subtle is readily evident from a simple example. Suppose that one were introduced to all the inhabitants of a village and asked to determine their relationships to each other. Obviously, possession of the same shade of hair would not indicate siblings; possession of slightly different shades of one color, cousins; and of different colors, members of unrelated families. But, just as obviously, hair color would be a datum essential for solution of the problem.

Every working taxonomist knows that some morphological differences do tend to be diagnostic of certain levels of classification, but the problem of determining this correspondence is essentially empirical and the values are properly assigned only *a posteriori*. Once so determined, there may be a degree of probability, not certainty, that similar values can be assigned *a priori* in studying allied forms. The basic data for the problem are extensive observations of the variation that can and does appear in defined groups. Many such observations are available, but many more are needed and most of those that are available need better analysis and presentation.

A major distinction is often made between quantitative and qualitative characters and examination of many diagnoses shows that subspecies and species are usually defined in quantitative, but genera and higher groups often in qualitative terms. This has been advanced as a rule, and there has even been argument as to whether "numerical" or "morphological" characters are better or more reliable in taxonomy (e. g., Ehrenberg, 1928). The distinction between quantitative and qualitative characters is real but by no means absolute. The difference is often a mere matter of convenience. For instance the absence of a premolar

in a mammal is customarily treated as qualitative, but the closely analogous absence of a scale in a reptile is often designated quantitatively. Many characters are treated qualitatively—e. g., strength or prominence of a crest on a tooth—simply because an easy means of measurement has not been devised, although the character may be more quantitative than qualitative in its real nature. With some exceptions it might be said that structures differ only quantitatively and that the only true qualitative differences are in the appearance of wholly new structures or the total loss of old. The practical distinction in study may be rather a matter of technique than of any real biological difference.

The characters of high taxonomic categories may show more obvious and constant distinctions than those of genera or smaller groups. The higher categories originate in ways that assure that such distinctions will exist genetically and the sharpness of the distinction arises mainly by slow accretion from species to species and by the disappearance of the transitional stages. On the level of genera close enough to each other not to represent different monotypic higher categories, this process is just beginning and the morphological distinctions may differ little in degree and not at all in kind from those that may appear between species, subspecies, or even individuals in one subspecies.

The fundamental point is not so much the distinctiveness of unit characters as their distribution in the groups; these groups, and not the characters, being the objects of classification. For instance, such a character as union or disunion of ectoloph and metaloph in a mammalian molar usually becomes widespread and relatively invariable within species, genera, or even families; hence, it is a character of those categories and is often felt to guarantee high categorical rank, but it may also vary within one interbreeding population and hence not be a so-called taxonomic character at all (Simpson, 1937a). Or, again, exactly the same morphological character may have different systematic value at different times in the history of one phylum. For instance, the most striking diagnostic character between the early Miocene horses, primitive species of *Parahippus*, and those of the preceding stages, *Miohippus* and advanced species of *Mesohippus*, is the uniform presence of a crochet on the upper molars of *Parahippus*. This is a generic character of that genus. But in transitional species of *Mesohippus-Miohippus* and in *Miohippus* the same character may appear as a variation within highly localized samples, purely individual variants of one subspecies, or may be a

variable but useful average character of a subspecies or a species (Schlaikjer, 1935; Stirton, 1940).*

If subspecies become species by isolation and species become genera by divergence and diversification, it is inevitable that diagnostic characters should thus appear as individual variations and tend gradually to become subspecific, specific, then generic characters, and that no particular kind of character should be characteristic of a particular taxonomic level. Paleontological examples of this process demonstrate that this sort of sequence does certainly occur at times and may be typical of evolutionary history. They thus demonstrate that geneticists of the cataclysmic school, like Goldschmidt (1940), are wrong for some cases, and therefore are wrong in general, because they claim that this process never occurs. There are, nevertheless, at least three frequent distinctions between intra-specific and extra-specific diagnostic characters. First, certain differences rise above the specific level less often than others; second, the accumulation of intra-specific differences normally gives rise to extra-specific differences of greater degree although not of different kind; and third, the whole number of significant differences is likely to be larger for higher categories than for lower.

As an example of the first sort, the most obvious of all animal characters, that of gross individual size, may be cited. Genera do usually differ in the size range covered by their species, just as species and subspecies usually differ in the size range of the included individuals. Thus size is frequently a generic as well as a specific character, but as a rule it is not a good diagnostic character for genera and so is not often used in defining them. An important technical reason for this is that genera (unless monotypic) do not have a true average size. From a given sample of a subspecies it is possible to observe and to measure a central tendency as regards size. Such a tendency does not necessarily exist for a genus and if it does exist it cannot be measured or estimated from one sample or in any other simple way. A more obvious reason is that related genera often tend to overlap very widely in size range, each with species more or less covering the optimum size ranges over a variety of local conditions, so that the sizes confined to one genus tend to be typical only of one or a few of its species and hence to be used for specific, not generic, diagnosis. This real but limited phenomenon is perhaps the origin of the too generalized dictum that quantitative characters are specific and qualitative, generic. Yet practical tribute to the fact that

*Of course it is possible to argue in the other direction and to say that an individual *Mesohippus* with a crochet belongs *ipso facto* to a different genus because a crochet is a generic character. This is a common tendency among practicing zoologists and is, indeed, what Schlaikjer did in this particular instance. For reasons expressed in this paper, I am convinced that this reasoning is fallacious.

genera do differ in this respect is paid by the zoologist who, when he finds a species far beyond the size range of known genera, immediately looks for, and usually finds, justification on other grounds for placing the aberrant-sized species in a new genus.

The quantitative increase of differences by increments, the number and hence the total of which is roughly proportional to taxonomic rank, is a factor known to and used by every zoologist. Its cause is related to such factors as sizes of mutations and rates of evolution. By experience, it is learned that differences of a certain degree cannot arise, or at any rate have not arisen in any known case, in a single step or within an interbreeding population. It is therefore justifiable to conclude that, when such differences are observed, they represent an accumulation of smaller differences such as accompanies divergence of specific or greater rank. Brachyodonty and hypsodonty in mammals provide a clear example. Relative height of cheek-tooth crowns varies in all groups, but only within narrow limits in one population. As far as has been determined *empirically (and up to now such determinations are necessarily empirical), a fully hypsodont tooth cannot arise from a brachyodont tooth in one or a few steps. Hypsodont and brachyodont* mammals never belong to the same genus. This implies not that height of crown is inherently a "generic character," but that the terms designate a high degree of difference that has generic or greater diagnostic value but is only the extreme on a continuous scale, on which lesser differences may be specific, subspecific, or individual. The generic degree is a simple sum of lesser degrees of difference.

The same element of rate of change is involved in the belief that adaptive characters are diagnostic of lower taxonomic groups than inadaptive, or habitus characters than heritage characters (e. g., Gregory, 1936). The higher taxonomic value of heritage characters is, in fact, a matter of definition rather than of any independent or esoteric biological relationship. Characters that have evolved slowly or not at all and that have been passed on from a more or less remote common ancestry to diverse descendents are by definition heritage characters and, also by definition, are characters of high taxonomic rank – indeed, "heritage" in this sense is merely another name for characters of high rank.

Adaptive characters are by no means confined to the diagnostic levels of low taxonomic groups. For instance, the streamlined contour of cetaceans is obviously adaptive and it characterizes a whole order (or a whole cohort). The point involved is not directly the adaptive nature of the character so much as its rate of evolution, in which the adaptive relationship to the environment is one of several important determinants.

If a potentially variable character is rather narrowly adapted to an environmental condition and if the environment changes or the group invades a new environment, then the character may also change relatively rapidly and differences in it may be diagnostic of minor units such as subspecies or still more localized groups. Pelage in protectively colored rodents (Dice, 1940) and other mammals is a typical example and such characters are largely responsible for the impression of zoologists that subspecies tend to differ in "superficial" ways of no higher taxonomic significance. If, however, the response of an adaptive character is more sluggish—for instance, from rarity of mutation, slight variation, or breeding structure crossing a very large population—a difference that does appear is likely to be of higher taxonomic value. The same will be true, as in cetacean body form, if the adaptation corresponds with a larger range of environmental conditions or if the pertinent environmental condition does not change.

The fact that higher taxonomic categories tend to differ in more characters than do lower categories is another aspect of much the same sort of evolutionary phenomena. Groups like subspecies that have some genetic transfer are not likely to differ markedly unless in the average conditions of characters of more immediate survival value. Once genetic transfer ceases and species become distinct, even pure chance tends to multiply the number of differences and this will become more extensive the longer the separation. This is a restatement of one of the reasons why morphological categories do tend to approximate genetic categories and why the foundation of a so-called phylogenetic classification on purely morphological data is justified.

NEOZOOLOGICAL AND PALEOZOOLOGICAL MATERIALS

The rest of this paper is devoted to some special problems and procedures more particularly related to the study of fossil vertebrates. Paleontology is as much a part of zoology as is the study of recent animals—a point here emphasized by using the names paleozoology and neozoology for the two major divisions of the subject. This subdivision arises from the different nature of available observational data, not from any fundamental difference in the aims of study or from any logical dissection of the science into "dead" and "living" parts. The principal differences in materials and data are as follows:

1.—Paleozoological specimens are all dead. So, in most cases, are the neozoological specimens used in taxonomy. Studies on the physiology and genetics of living animals have been made for such a very small number of species that they do not constitute the data of neozoological

taxonomy but are only examples useful in interpreting data derived from dead specimens. They serve the same purpose for both paleozoological and neozoological taxonomists.

2.—With unimportant exceptions, paleozoological specimens comprise only bones and teeth, and generally only a fraction of these for each individual. Neozoological data could include the whole anatomy of each animal, but they seldom do in practice. At least through the generic level, neozoologists base classification almost exclusively on external characters. Ichthyologists and herpetologists do, as a rule, collect whole animals, but rarely use internal characters on these levels of classification. Ornithologists collect only skins for ordinary taxonomic purposes. Mammalogists collect skins and skulls. The use of mainly external characters by neozoologists and of exclusively internal characters by paleozoologists is the most striking difference in their materials.

3.—Neozoologists sometimes have larger samples than paleozoologists. The difference is not as great as might be supposed. There is no absolute criterion as to what constitutes an adequate sample for taxonomic purposes, but I would judge that samples of more than ten specimens will usually suffice to establish the reality and basic distinctions of a subspecies, if efficiently used and analyzed. More than fifty specimens, analyzed with equal efficiency, will usually permit a virtually complete definition. Such definition is exceptional both in neozoology and in paleozoology. As a fairly typical example, in a recent neomammalogical revision competently based on all the pertinent materials of several great museums (G. M. Allen, 1938, 1940), the samples studied by the author were inadequate (ten or fewer specimens) for 58 per cent of the unit groups (subspecies and monotypic species), adequate but incomplete (eleven to fifty specimens) for 31 per cent, and fully sufficient (over fifty specimens) for only 10 per cent. There does not happen to be a closely comparable recent paleomammalogical revision giving such data as to sizes of samples, but a census of materials for a characteristic Tertiary fauna (Lebo) shows: inadequate, 65 per cent; adequate, 31 per cent; fully sufficient, 4 per cent. Collections for some of the well-known later Tertiary faunas have fewer inadequate unit samples and those for some recently discovered or poorly collected faunas have more. In general it is evident that paleomammalogists have about the same percentage of adequate samples but have fewer large, relatively complete samples.* In both fields the sizes of available samples are steadily increasing. For

*It is only fair to add that the "classical" methods of mammalogy are so inefficient that they do not, as a rule, obtain any more information from samples of two or three hundred specimens than could be obtained by efficient use of twenty or thirty. Thus the neomammalogists have derived little real advantage from their large samples. Of course there will be improvement in this respect and the large samples are available for it.

some of the groups other than mammals, students of fossils are at a greater disadvantage. For instance there are only one or two species of dinosaurs for which adequate samples have been collected. For this reason, if for no other, the specific taxonomy of such groups is highly unreliable or practically non-existent.

4.—Neozoological samples can and frequently do cover the whole areal range of the included groups and the whole number of distinct groups present in a given area. Paleozoological samples practically never do either. This is much the most important permanent disability of paleozoology. It means that paleozoology can contribute relatively little to some taxonomic problems, especially those having to do with geographic variation and horizontal subspecies.

5.—Paleozoological samples can cover long periods of time. This is the great advantage of paleozoological data. The neozoologist's universe has no time dimension. For all practical purposes his subjects are all contemporaneous. If the time observable suffices for any significant change in his populations, this change does not proceed beyond the lowest levels of subspecies or still smaller groups. He may thus take the discontinuities between species, genera, and higher units as absolute, a fact that greatly simplifies the taxonomic task but that removes from his direct observation a fundamental goal of research: the mode of origin of these discontinuities. The paleozoologist, on the other hand, is as much concerned with temporal as with geographic sequences. He is so constantly dealing with time that his whole manner of thought is affected by it. The added dimension greatly complicates his practical task of classification, but at the same time gives him a direct approach to major evolutionary patterns and processes.

These differences in materials for study not only necessitate differences in procedures of classification but also induce differences in attitudes toward classification and toward the broader problems of taxonomy. They emphasize the desirability of an understanding of both neozoological and paleozoological contributions to these problems and the absolute necessity of a synthesis of the two fields for further progress toward their solution.

With differences caused by personal taste and knowledge, all zoologists recognize the same classes, orders, and families. These can all be defined either in paleozoological or in neozoological terms and a family of fossils is the same sort of thing as a family of recent animals. When it comes to the levels being discussed in this symposium, genera, species, and subspecies, this equivalence can no longer be taken for granted. As regards these low taxonomic ranks there is real room for question whether a

single, unified zoological system is an attainable ideal or whether paleozoological and neozoological classifications must forever be different. The problem of vertical units enters into this, but it is distinct and will be discussed separately. The present question is whether the horizontal genera, species, and subspecies of paleozoology and neozoology are or can be units of about the same real taxonomic rank. The question is an old one. More than a century ago Hitchock (1836) wrote: "When I speak of species here, I mean species in oryctology [paleontology], not in ornithology. And I doubt not, that in perhaps every instance, what I call a species in the former science, would be a genus in the latter." Zoologists are still wondering whether this may not be true.

Genera

For genera, it is possible to give a definite answer: paleogenera and neogenera (if I may be permitted a self-explanatory barbarism) can be exactly equivalent and usually are approximately so. I have yet to see a genus of recent mammals, or, at least, one that had any good chance of being valid and having this rank in a reasonable classification, that could not be recognized from a single specimen of the skull. Most of them can be recognized from a single jaw, and many from a single tooth. Thus neither incompleteness of individual specimens nor small size of samples prevents a paleozoologist from recognizing neozoological mammalian genera. Studies of Pleistocene faunas, in which neozoological genera occur as fossils, provide an experimental check on this and confirm the equivalence and recognizability. Among some of the lower vertebrates identification of recent genera sometimes requires more complete knowledge of the skeleton, but even in these groups the usual paleozoological data generally seem to suffice for this purpose. Paleozoologists frequently discover that they can recognize genera from skeletal characters that are not used by (and are often unknown to) the neozoologists who founded the genera. For instance, neornithologists never define genera on osteological characters, but paleornithologists have no serious difficulty in referring Pleistocene bird bones to the proper recent genus (e. g., Howard, 1930).

It is, nevertheless, probable that paleozoologists tend to draw generic lines somewhat more broadly than do neozoologists, although in both fields the tendency is obscured by great differences between individual workers. If real, the tendency is not a matter of necessity, and certainly not a matter of erroneous interpretation, but one of taste. The paleozoologist has, on an average, fewer species and subspecies within a genus of given scope but has a greater variety of generic and, especially, higher

groups and is much concerned with clear expressions of relationship on upper taxonomic levels. The use of rather inclusive genera is therefore more convenient and natural for him. The neozoologist has greater numbers of subspecies in a given morphological range and is mainly concerned with the arrangement of these and other low taxonomic groups, so he tends to use smaller and smaller genera. For instance, the great number and small scope of genera of recent murid and cricetid rodents commonly recognized by neozoologists seem to the paleozoologist unjustified by the morphological (or probable genetic) facts.

My own sympathies in this battle of the lumpers and the splitters naturally incline toward the broader, paleozoological sort of inclusive genera, but a compromise is possible and may eventually prove acceptable to both. I am informed (by Dr. Mayr) that the ornithologists, having once broken all their genera into small fragments, are now engaged in putting them together again. Perhaps the neomammalogists will soon enter this new lumping phase. As for the paleomammalogists, they may have the splitting phase still to suffer, but may happily be spared its most extreme form.

Species

For many years paleozoologists have gone on naming hundreds and thousands of species, some of them never doubting that these were true species in the neozoological sense and most of the rest not caring whether they were or not. A note of doubt has been heard from time to time, more frequently in recent years. Some of the best paleozoological taxonomists (e. g., Jepsen, 1933; Scott and Jepsen, 1936) have even suggested that it may not be possible to recognize true species (in the neozoological sense) among fossil vertebrates, at least for the present.

There are really two questions. Are paleontological species real groups? Are they approximately equivalent to taxonomic species as defined on a previous page from a more neozoological point of view? Like neozoologists, and with more excuse, paleozoologists have created an enormous number of synonyms and have also given many names to what may have been but cannot be shown to be real and newly discovered groups. Fossil species are usually first described from fragments, sometimes not homologous with the known parts of related forms, and many supposed species have been based originally on single specimens. These are inevitable results of the nature and history of paleontological discovery. Earlier paleontologists had no real idea of the extent of morphological variation that can occur in a single species and workable criteria have only slowly been achieved, hand in hand with similar work by

neozoologists and with experimental work. It is conservative to guess that among previously proposed species of fossil vertebrates, aside from types of currently recognized genera, not more than a quarter represent natural and distinct groups. The fraction of valid species is probably much lower.

This situation is certainly deplorable, but it exists to greater or less degree in every field of zoology. It can be and is being cleared up by revision with enlarged samples and by the use of more rational and objective criteria for intraspecific variation, especially those based on the statistical relationships of samples and populations. No matter how many mistakes may have been made in the attempt to do so, there is little real doubt that the groups called species in competent recent paleozoological studies tend to correspond with real units of population in nature.

Strict application of competent modern methods to adequate samples should seldom or never result in the recognition of false groups (aside from the element of human error by the student), but it may fail to distinguish two groups properly of specific rank and it may result in calling groups species when their true rank may be higher, e. g., genera, or lower, e. g., subspecies. The extent to which this may result from peculiarities special to paleontology is a measure of the degree in which the species of paleozoological taxonomy are likely to differ from those of neozoological taxonomy.

The fact that a neozoologist could and that a paleozoologist could not have recourse to actual breeding experiments in a critical case has little practical significance at present. How many vertebrate species have, in fact, been defined experimentally in a way importantly different from previous morphological definitions? All will admit that the number is insignificantly small. Larger size of unit samples for recent animals is also a minor consideration. Although it does impede the paleozoologist in certain particular cases, it is neither a general nor a necessary disability. The better spatial distribution of neozoological samples is more pertinent to the recognition of subspecies, but as regards species it may tend to have an effect opposite to that usually supposed: it is likely to make the paleozoological species a smaller, not larger, unit than the neozoological species. Terminal groups that are shown by intervening samples to be subspecies of a single species may frequently be mistaken for species if the intervening samples are lacking, which is more likely to be true in paleozoology than in neozoology.

The important difference between paleozoological and neozoological taxonomic species—if such difference does necessarily exist—must arise

from the availability and use of different morphological characters. It is a common belief, on both sides, that paleozoologists cannot recognize neozoological species because the latter are defined by external characters. This statement contains a glaring (but sometimes unnoticed) fallacy: it rests on unstated and unproved postulates. These postulates are that external and internal characters are fundamentally different in taxonomic value, that their variations do not tend to be closely correlated, and that their population distributions are significantly dissimilar. Making these assumptions by implication and without conscious analysis is unscientific. They demand more careful statement and study than has yet been given them.

There is, as far as I know, no evidence that external and internal characters have any different genetic basis or involve any different hereditary processes. On the contrary, all the pertinent data suggest that they are exactly analogous in this respect. Can two organisms differ only in external, or only in internal, characters? Since it is theoretically possible for organisms to differ only in a single character, the answer to this question is evidently "yes," but the question does not mean very much for taxonomy. Taxonomy is more pragmatic and is not concerned so much with what can as with what does happen. Distinct populations that differ in only one or two characters are so unusual as to have almost no bearing on practical taxonomy. Even a single difference in alleles commonly results in more than one phenotypic "unit character" difference. In nature most distinct populations differ in average values of dozens or of hundreds of morphological peculiarities.

The chances that all such multiple differences will be either external or internal can be worked out for various sets of permissible postulates. For instance, if only ten differences are involved and these are as likely to be internal as external, there is less than one chance in one thousand that all will fall into a particular one of these categories From such considerations, it is easily shown that the chances of all such differences being external are quite negligible unless the differences are very few in number, unless there are many more distinguishable external than internal characters, or unless external characters are inherently much more likely to vary than internal characters. It is an established matter of experience that such conditions, if they ever obtain for the characters available for distinction of species, are certainly highly unusual.

Regardless of these factors related to the genetic incidence of differential morphological characters, it could still happen that the taxonomic significance of external and internal characters was different. First, the segregation of the two could conceivably be different, so that populations

defined in terms of one would have distinctly different boundaries from those defined in terms of the other even though comparable in scope. This possibility need not be considered in detail because it follows from the similar hereditary determination of the two and from the basic definition of species in terms of breeding structure that no consistent tendency toward such a difference can exist, even though some difference of the sort might occasionally arise by chance and at random Second, it is conceivable that some non-genetic factor, notably natural selection, might operate so much more strongly on one than on the other that populations defined in terms of the first would be more sharply delimited and smaller in scope than those defined in terms of the second

This is the most essential point involved in the present problem. It deserves a great deal more attention than it has ever received and more detailed discussion than can be included here Data for checking it empirically and objectively are scanty, but some are available TABLE 1

TABLE 1

VARIATIONS OF CHARACTERS IN *Peromyscus*

A THREE TYPICAL "INTERNAL" CHARACTERS

Stock	Linear Bone Dimensions					
	Mandible		Condyle-Premaxilla		Bullar Width	
	M	*SR*	*M*	*SR*	*M*	*SR*
Alexander, Iowa	15 36	14 04–16 68	22 81	20 50–25 12	10 26	9 17–11 35
Moville, Iowa	17 39	16 05–18 73	25 62	23 79–27 45	11 25	10 47–12 03
Greenland, N. H	16 72	15 42 18 02	25 05	23 24–26 86	11 02	8 45–13 59
Vineyard Haven, Mass	17 72	16 43-19 01	26 26	23 95–28 57	11 40	10 71–12 09

B TYPICAL "EXTERNAL" CHARACTER

Stock	Tint Photometer Readings on Dorsal Stripe					
	Red		Yellow		Green	
	M	*SR*	*M*	*SR*	*M*	*SR*
Alexander, Iowa	5 10	1 21–8 99	4 25	1 01–7 46	3 67	1 08–6 29
Moville, Iowa	8 59	3.43-13 75	7 07	2 57–11 57	5 75	1 44–10 06
Greenland, N H	8 65	5 63–11 07	7 44	5 02–9 86	6 36	3 94–8 78
Vineyard Haven, Mass	8 03	3 13-12 93	6 67	2 55-10 79	5 27	2 44–8 10

M is the mean, as given by Dice and *SR* is standard range from standard deviation (see Sim 1941), calculated from Dice's data, and here expressed by limits rather than by span

is a good example, based on Dice's observations (1937a, b, 1939) on uniformly laboratory bred stocks of *Peromyscus* from different localities.

Although it is unlikely that there is any genetic or any selective correlation between skull dimensions and color of pelage, it is evident that the inferences as to resemblances and differences of these four populations would be the same whether based on the internal or on the external characters. On the basis of either sort of character, the two geographically close stocks from Iowa differ more than do the three geographically scattered Moville, Greenland, and Vineyard Haven stocks. The Moville and Greenland stock are particularly similar, almost indistinguishable, and the Vineyard Haven stock is a little more distinctive. The mammalogists' taxonomic expression of these facts is to place the Alexander population in *Peromyscus maniculatus* and the other three in a different species, *Peromyscus leucopus.* The Moville and Greenland populations are local variants of one subspecies, *Peromyscus leucopus noveboracensis,* and the Vineyard Haven group has been placed in a different subspecies, *P. leucopus fusus* (although it may be noted that this arrangement rests more on the relatively sudden transition to adjacent *P. l. noveboracensis* than on the degree of difference—a point in subspecific definition not particularly pertinent at this point in the discussion).

Data on other variates of these samples, on many other samples of allied and intervening populations, and on a wide variety of other mammals support the same generalization natural populations of mammals tend to differ about as much and about as sharply in skeletal characters, such as dimensions of skull and teeth, as in external characters, such as pelage color, and the groups based on the two sorts of characters by analogous methods of inference tend closely to coincide. There are a few real exceptions to this generalization and there are supposed exceptions that are more apparent than real. Among the latter may be counted such classic cases as the lion and tiger or the horse and zebra, so obviously different when seen in the flesh and so apparently similar when seen as skeletons. But this is only a matter of ease of observation. The species are really distinct in osteology also, when the skeletal characters are closely and correctly analyzed, and the osteological species are the same in scope as the pelage species.

It may be concluded that paleozoologists can, as a rule, recognize the same sort of species as do neozoologists. It is undoubtedly true that they frequently do not do so. This is due in part to personal factors—certainly not all neozoologists recognize the same species; in part to the fact that specific taxonomy is a retarded study in paleontology, only now

slowly emerging from the days of rule-of-thumb and excessive subjectivity; and in part to the paucity of well-distributed paleozoological samples. I believe that the tendency in the two sciences is for their specific concepts to become more and more similar, although this question of sample distribution does involve an important difference more explicitly stated at the end of the following discussion of subspecies.

Subspecies

Few vertebrate paleontologists ever propose subspecies and those who use them at all do so sparingly. Checking the last 25 papers on fossil mammals to reach my desk, I find that 24 of them do not so much as mention subspecies, although they include some large faunal revisions based on excellent collections, and that 49 new species are proposed and only 2 subspecies. In contrast, 25 recent taxonomic papers on living mammals, taken at random, propose 6 species and 6 subspecies. The disproportion would have been greater except for the accident that one of these papers is by one of the few students who habitually proposes subspecific names for fossils. It is also significant that both the supposed new subspecies that are described are recorded as occurring in direct association with much more abundant remains of the typical subspecies of the same species. They receive separate designation only because they seemed to the author too far from the average condition to be called typical, the criteria used being entirely subjective as far as shown. The chances are great that these are not subspecies at all but are either artificial groups based on individual variants within the typical groups or are distinct species.

The number of true subspecies in the paleontological literature, that is, of subspecies that are really analogous in structure and scope to those of sound neozoology, is certainly extremely small. Practically speaking, it is not too much to say that paleozoology does not treat with subspecies in this sense. We may briefly discuss why this is true now and whether it is necessarily and permanently true.

From the considerations summarized in the preceding discussion of species, it is fairly well established that paleozoologists could recognize, and would tend to arrive at, the same sort of subspecies as neozoologists if they had samples analogous in size and distribution and if they treated them in a similar way. Neither of these conditions has been commonly satisfied in the past.

The very fact that their materials usually consisted of a few isolated specimens during the early and classical periods of their science meant that paleozoologists perforce developed the attitude and methods of com-

parison of individual specimens. Their development of group concepts and of the methods of population inference from samples has been retarded; indeed even now many active paleozoologists hardly understand what these words imply. In some respects the higher the taxonomic group, the less subtle and difficult are the methods of group inference. Moreover work on the higher groups seems more important and it is a field that neozoologists have recently tended to underemphasize. Thus paleozoologists have not made any serious effort to study such small groups as subspecies, partly because they did not know how and partly because they did not want to spend time on what they felt to be a relatively unimportant subject.

These differences of method and aim are mainly historical and psychological and are not likely to persist except as they are forced by necessity. The fact is, however, that paleozoologists do not have the materials for recognition of the horizontal subspecies of most of their species and that they are not likely ever to have them except in a minority of special cases. Such subspecies can only be truly defined on the basis of a considerable sequence of adequate, contemporaneous samples scattered over most of the range of the species. Paleozoologists have few sample sequences of this sort (with an increasing number of exceptions mostly in the Pleistocene), and it follows from the conditions of deposition and preservation of fossils that they will never have adequate sequences for a majority of their species.

They do have now a few and certainly will have many more series of scattered samples that cover not the whole areal and ecological range of a species but at least the range of more than one subspecies. These permit and will eventually necessitate the use of the subspecific category in paleozoological taxonomy, work that certainly can be but has not yet been properly done. Even in such cases, however, it is quite impossible to guarantee the contemporaneity of dispersed samples within a few millenniums. This introduces temporal as well as spatial variation and hence there is an element involved in the relationship of taxonomic to phylogenetic paleozoological subspecies that is necessarily somewhat different in kind from this relationship in neozoological subspecies.

It results from the sampling conditions that a majority of fossil species, as they are known and definable, are essentially monotypic for any one time, that only one subspecies is usually represented by the available demonstrably contemporaneous samples of a species. It follows, with apparent paradox, that the paleozoologist, who rarely mentions subspecies, deals almost exclusively with subspecies rather than truly dealing with species. The group that he can and does really envision from his

samples is usually a subspecies, but since no other subspecies of the same species is usually known to him, this inferred subspecific group is taken to represent a species. It is, in fact, the species of most paleozoological taxonomy. Since it is probable that most extinct species did have two or more subspecies in nature, this involves a systematic difference between paleozoological and neozoological taxonomy. Having only one subspecies in each species (as a rule), the paleozoologist has no occasion to distinguish subspecies of one species, but only to distinguish subspecies each of a different species. Thus he may be and usually is defining subspecies, although this has so little evident importance or practical significance that it is not even noticed, and he is diagnosing (not really defining) species.

VERTICAL UNITS

The great, special problem of paleozoological taxonomy is the definition and subdivision of units that have an extension in time as well as in space. If species arose discontinuously, as claimed by Goldschmidt and a few others, the problem would not exist, or at least it would be no different from the recognition of the *de facto* discontinuous species of horizontal classification. In such a case, there would be a definite point in time when each new species arose and the specific boundaries would be real and visible, as much so as if species were invariable units resulting from divine creation. Then subspecies would also be without gradations into any units except other subspecies and, with sufficient samples in hand, should be absolutely definable as single phylogenetic branches even though they changed slightly in time. Although somewhat less clear-cut, genera, too, would be rather easily definable; they might not arise full-blown at one step, but could at least be separated at a sharp interspecific boundary.

From a practical point of view this pleasingly simple situation does actually exist for a considerable part of paleozoological taxonomy, whether or not this theoretical reason for it be accepted. A large number of species and genera do appear suddenly, without closely similar predecessors, in the paleontological record and do disappear just as sharply, without known immediate descendents. As long as this condition of the evidence exists, classification is simple and requires no special consideration of the time dimension, whether the condition is supposed to arise from the non-existence or from the non-discovery of intermediate stages. A century ago, this was so universally the case with known species that paleontology was regularly cited as evidence against the theory of organic evolution. The apparent sudden origin of species is no longer universally

true but it is still true often enough to be cited—not, as a rule, by paleontologists—as evidence against the continuity of morphological evolution.

Whatever may be held as to the causes of the breaks or as to the universality of continuity, there are now many known examples of longer or shorter sequences in which there is no definite and real discontinuity, each population more or less overlapping in variation those preceding and those following it, but which change so much that every taxonomist places the earliest and latest members of the line in different species or genera. Clearly a species as a subdivision of such a temporal, or vertical, succession is quite a different thing from a species as a spatial, or horizontal, unit and cannot be defined in the same way. The difference is so great and, to a thoughtful paleozoologist, so obvious that it is proper to doubt whether such subdivisions should be called species and whether vertical classification should not proceed on an entirely different plan from the basically and historically horizontal Linnaean system.

In line with these theoretical or philosophical misgivings, various distinct terms have been proposed for successive stages within a single vertical line. Of these "mutation," as proposed by Waagen (1869), is most important and it alone has been very widely used. Unfortunately the geneticists, most of whom seem to be unaware that this is the prior and historically correct use of the word "mutation," have used the same term with a sharply different and yet obliquely related meaning. A great deal of misunderstanding and of mutual irritation between geneticists and paleozoologists has resulted. Now the paleozoologists may just as well abandon their priority and admit that the geneticists have carried the day and have succeeded in purloining the word. We have no more chance of retrieving it than the Indians have of retrieving Manhattan and the sooner we abandon it completely, the sooner we will be able to express ourselves intelligibly to the geneticists, and in turn to understand them without unnecessary difficulty

So far none of the varied proposals for non-Linnaean arrangement and nomenclature of vertical units and their successive subdivisions has been widely accepted and none seems promising at present This is not the place to go into detail as regards the reasons for this failure or the serious and intricate difficulties of the problem, and it is emphatically not the occasion for making any new proposal of this kind.

We do, in fact, use the same system of nomenclature for subdivisions of a vertical line as for subdivisions of a horizontal distribution. The tendency of paleozoological practice is this: *successive taxonomic units are inferences as to morphological units such that the net difference in morphology between corresponding parts of those units is of the same order as*

that between horizontal units of the same rank in the same or allied groups. For instance, in the main line of horse evolution, the average difference in structure between successive genera tends to be of about the same order as the average difference between closely allied but distinct contemporaneous genera of perissodactyls. No claim can be made that this practice is perfect or even that it is theoretically desirable, but it is what paleozoologists really tend to do, it works fairly well for them, and no one has yet proposed a system generally believed to be more promising.

The line between such units within essentially continuous sequences must, of course, be arbitrary. Although the average difference between successive genera or the difference between their medial or central (not necessarily typical in a technical sense) species is comparable to that between contemporaneous genera, the difference between the last population placed in one genus and the first placed in the next may be, and in the postulated circumstances must be, comparable to that between contemporaneous subspecies or still more nearly related groups. Although specialists usually seem to reach some sort of working agreement as to the most convenient approximate position for these artificial boundaries, their exact position is determined by no general rule or criterion and naturally remains subject to doubt and dispute.

In practice the boundary is usually placed in one of two ways. Quite commonly, the sequence in question was not really continuous when its successive units were first defined—that is, knowledge of it was not continuous. The gaps in knowledge were then used to separate the units. When the series is filled in, the boundaries are inevitably placed somewhere within what were the gaps. If the accidents of discovery had first revealed other points along the line, the eventual subdivisions of the completed continuous sequence would have been in quite different positions on it. If the gap was large, there is a strong tendency for the worker whose material fills the gap to define a new unit rather than to refer his novelties to one or the other of the adjacent units, even when such a step is not warranted by the pragmatic rule of comparable morphological scope in units of the same rank. This human tendency to seek the solution of problems by evading them makes more difficulty in the end because eventually it only means that two boundaries have to be settled where only one existed before, or was necessary. Hence arises much of the needless splitting and complication in paleozoological nomenclature.

The second and somewhat more rational but not always more practical way of arriving at such boundaries is to select some feature or features of essential importance characterizing a genus or species and to draw the

line where this character becomes dominant or universal in the evolving population. The character may be selected because it is easily observed and helpful to a hurried taxonomist in his capacity as sorter and cataloguer of specimens, e. g., the crochet in *Parahippus*. Or it may be a character of primary selective value in the economy of the animals concerned, e. g., cement in *Merychippus*. Usually both factors operate in varying degree. Even such characters do not arise all at once and they still define a region rather than a point in the sequence, but if they are well-chosen they certainly provide the best possible basis for such subdivision. This sort of criterion is not, however, available in all cases, or even in most. It generally involves the appearance of a new structure or character and this is a rare event in evolution compared with the gradual modification of an existing structure. Thus the criterion can seldom be used to distinguish successive species, which are much more numerous than are the new structures involved in their evolution—another reason why so-called qualitative characters are commonly believed to have inherent superspecific rank.

Seen in narrower perspective, the taxonomically troublesome continuous vertical sequences of paleozoology are closely analogous to the horizontal sequences of neozoology called clines by Huxley (1938). Although no new principle or discovery is involved in this restatement of known facts, the terms of this restatement do involve a new way of looking at the facts and this will, I think, prove to have far-reaching and even fundamental effects on zoological theory. Sufficient study of this aspect of the matter requires considerable detail and I have in hand a special paper on it that I plan to present elsewhere. It is, however, sufficiently pertinent to the present subject to demand mention here and to warrant the inclusion of one illuminating example.

As originally proposed, the idea of a continuous cline is that of a sequence of contemporaneous populations arrayed geographically and intergrading in progressively changing morphological characters. Although many such sequences probably also show some gradation that is solely phenotypic, the gradation is normally maintained by genetic exchange, actual interbreeding of adjacent segments of the population. Now if a sequence of successive populations is arrayed temporally, it normally shows a similar sort of continuous intergradation in progressively changing morphological characters. The cause is different (although not unrelated), but the objective sequences are so precisely analogous that I see no reason why these may not usefully be given the same name: clines. Clines may, then, be distinguished according to the variate that is used to define the array. In one case the arrangement is

geographical and these may be called choroclines, i. e. "space clines." In the other, the arrangement is temporal, and these may be called chronoclines, i. e. "time clines."

The following example illustrates a typical chronocline and serves to point the contrast between defining a vertical chronocline, with various successive horizontal subdivisions, and defining "vertical species" in the usual sense. Data are given for only a single variate, but other variates in the same samples confirm the conclusion, some more and some less clearly. The named stratigraphic subdivisions from which the samples come are arranged in temporal sequence from left, Clark Fork, oldest, to right, Lost Cabin, youngest. *Ectocion* is a genus of primitive ungulates, condylarths, typical of this part of the North American faunal sequence. (See also Simpson, 1937b.)

The "vertical species" of the table represent routine identification by

TABLE 2

Ectocion IN THE UPPER PALEOCENE AND LOWER EOCENE OF WYOMING
AMERICAN MUSEUM SAMPLES

Length of M_1 in mm.	Frequency in Each Horizon				So-Called "Vertical Species"
	Clark Fork	Sand Coulee	Gray Bull	Lost Cabin	
5.3–5.6	1*	0	0	0	*E. parvus*
5.7–6.0 6.1–6.4	1 6*†	0 3	0 1	0 0	*E. ralstonensis*
6.5–6.8 6.9–7.2 7.3–7.6	4 1 0	4† 3 0	7 2 2*†	0 0 0	*E. osbornianus*
7.7–8.0 8.1–8.4	0 0	0 0	1 0	0 1*†	*E. superstes*
Ascending stages in unit phylum or chronocline.	*E. osbornianus ralstonensis*	*E. osbornianus complens*‡	*E. osbornianus osbornianus*	*E. osbornianus superstes*	True Shifting Vertical Species or Chronocline *Ectocion osbornianus*

*Type of "vertical species."
†Type of subspecific stage
‡Completion of the example unfortunately requires the proposal of a new name at this time **Ectocion osbornianus complens.** Type Amer. Mus. No. 22498. Hypodigm American Museum specimens of *Ectocion* from the Sand Coulee of Granger. Diagnosis Ranges of all characters overlapping those of *E. o. ralstonensis* and *E. o. osbornianus* but means intermediate between those two groups, length of M_1 (ten specimens) 6.72 ± 0.11.

my predecessors at the American Museum and by me in earlier years. They follow what was, and still is to a large extent, standard practice in paleontology. Such practice is supposed to produce a classification in which the vertical, time element is taken into consideration, because the species are commonly given such a dimension and shown as running vertically through various horizons. In this case, and innumerable others similar in character, this supposition now seems to me naive. Study of the distributions as a whole shows beyond much doubt that the "species" of this system are purely subjective size groups. They do not tend to correspond with any real, defined populations that existed in nature and therefore they are not really species in any sense of the word. Instead of taking the effect of the time dimension into consideration, they ignore and conceal it.

Although it is always possible that some extraneous specimen has crept in, what seems really to have happened in this case is that the sample from each horizon was derived from an essentially homogeneous population. Each of these (more or less) contemporaneous populations appears to have been derived from that preceding it in the same general area and to have given rise to that following it. At any one time there was then only one species, apparently only one subspecies, and the genetically continuous, ancestral-descendant, series of populations gradually changed in morphology, most noticeably in size as shown by the exemplifying data of the table. The resulting picture is typically that of a cline, and in this case a chronocline.

Such a cline as a whole might be given any taxonomic rank in the Linnaean system, depending on how great is the morphological difference between known early and late members. In the example, my judgment is that the chronocline is of specific scope. This is supported by the fact that all the ranges of the unit populations overlapped. (The samples do not overlap in every character, but some of the samples especially that from the Lost Cabin, are very small and the populations almost certainly did overlap.) Even in Lost Cabin times there were undoubtedly some Ectocions of this lineage that were as small as some of the larger individual variants in the long precedent Clark Fork and that could indeed hardly be distinguished from the latter if they were compared as individuals.*

*Here is a point that is a subject for criticism, not to say ridicule, among some paleontologists of the old school. If two specimens cannot be distinguished, except by trivial variations admittedly less than can occur within a subspecies, how can one maintain that they are taxonomically distinct? If a fossil can only be identified when its horizon is known, what becomes of the whole basis of paleontological correlation of horizons by the identification of their fossils? Both objections are based on the fallacious tendency to compare *individuals* when the correct comparison is of *groups*. The groups as such are here readily distinguishable even though some individuals are not. Valid inference of group characters requires some homogeneity as to time, and hence some specification as to horizon.

The chronocline is thus essentially continuous and is definable as a species, for which in this example the valid name is *Ectocion osbornianus*. This natural chronocline species is of course quite different in character and limits from the wholly artificial "vertical species" hitherto given that name. Since the species consisted of morphologically different populations at different times, it is both theoretically and practically valuable to have some means of subdividing it and of designating different stages in its development. It appears that each of the samples, sorted by collector's specifications, is distinguishable from the others in average characters, so that the number and character of the chronocline subdivisions are automatically determined by the stratigraphic subdivisions recognized and recorded by the collector. If he had used different stratigraphic units, the taxonomic units would also be different. The lines drawn between the different taxonomic subdivisions are in this sense arbitrary. It does not follow that the taxonomic groups are artificial or unreal: they are natural groups approximating populations that once existed in nature. In this they differ profoundly from the completely artificial "species" of the old classification.

Since the chronocline has been designated by a Linnaean binomial, its subdivisions may conveniently be designated by trinomials. This is also justified by the fact that their average scope, resemblances, and differences are fairly analogous to those of subspecies along a chorocline in neozoology. Although paleontologists only very exceptionally have materials permitting the recognition of subspecies such as enter into choroclines and are recognized by neozoologists, they frequently have materials fully adequate, when carefully analyzed by group methods, for the recognition of analogous units in chronoclines.*

In studying choroclines, some are found to have almost even slope (in graphic terms) from one end to the other, while some have distinct plateaus bounded by shorter steep slopes or narrow transition zones. Ideally, it is the latter phenomenon that permits the definition of well-defined and homogeneous subspecies. There is considerable evidence that a similar phenomenon occurs in chronoclines on a much larger scale. The chronocline analogue of the steep transition zone in a chorocline is a relatively brief period of relatively rapid evolution. Even within an essentially continuous sequence, an acceleration of this sort provides a definite and natural boundary zone (although not a line or point). For many reasons too complex to list here, it is clear that these zones of acceleration are least likely to be represented by fossils, and are almost sure

*And also, still more commonly, they have low-rank units, subspecific in morphological scope and distinctiveness, in which both time and space are a factor. This is a complication that need not be considered in the present summary.

to be more poorly represented than are the periods of more even and slower evolution. Thus it happens that the inevitable and at first sight merely accidental division of vertical units by gaps in the record does probably tend to approximate a real and important sort of division in phylogeny. This phenomenon apparently has little bearing on the lower levels of taxonomy but it has probably been active in relation to some genera and may usually be involved in the delimitation of higher taxonomic categories.

LITERATURE CITED

Allen, Glover M.

1938, 1940. The mammals of China and Mongolia. Natural History of Central Asia **11**: Pts. 1-2. Amer. Mus. Nat. Hist.

Allen, J. A.

1919. Severtzow's classification of the Felidae. Bull. Amer. Mus. Nat. Hist. **41**: 335-340.

Crampton, H. E.

1932. Studies on the variation, distribution, and evolution of the genus *Partula*. The species inhabiting Moorea. Pub. Carnegie Inst. Washington, No. **410**.

Dice, Lee R.

1937a. Additional data on variation in the prairie deer-mouse, *Peromyscus maniculatus bairdii*. Occas. Pap Mus. Zool. Univ. Michigan, No. **351**.

1937b. Variation in the wood-mouse *Peromyscus leucopus noveboracensis*, in the northeastern United States. Occas. Pap. Mus. Zool. Univ. Michigan, No. **352**.

1939. Variation in the wood-mouse, *Peromyscus leucopus*, from several localities in New England and Nova Scotia. Cont. Lab. Vert. Genet. Univ. Michigan, No. **9**.

1940. Intergradation between two subspecies of deer-mouse (*Peromyscus maniculatus*) across North Dakota. Cont. Lab. Vert. Genet. Univ. Michigan, No. **13**.

Dobzhansky, Theodosius

1941. Genetics and the origin of species. 2nd ed. New York, Columbia Univ. Press.

Ehrenberg, K.

1928. Betrachtungen über den wert variations-statistischer Untersuchungen in der Paläozoologie nebst einigen Bemerkungen über eiszeitliche Bären. Pal. Zeits. **10**: 235-257.

Goldschmidt, R.

1940. The material basis of evolution. Yale University Press.

Gregory, William K.

1936. Habitus factors in the skeleton of fossil and recent mammals. Proc. Amer. Phil. Soc. **76**: 429-444.

Hitchcock, Edward

1836. Ornithichnology. Description of the footmarks of birds (*Ornithichnites*) in New Red sandstone in Massachusetts. Amer. Jour. Sci. (1) **29**: 307-340.

Howard, Hildegarde

1930. A census of the Pleistocene birds of Rancho La Brea from the collections of the Los Angeles Museum. Condor **32**: 81–88.

Huxley, Julian S.

1938. Species formation and geographical isolation. Proc. Linnaean Soc. London **1937-38**: 253–264.

Jepsen, G. L.

1933. American eusmiloid sabre-tooth cats of the Oligocene epoch. Proc. Amer. Phil. Soc. **72**: 355–369.

Mayr, Ernst

1940. Speciation phenomena in birds. Amer. Nat. **74**: 249–278.

Pocock, R. I.

1917. The classification of existing Felidae. Ann. Mag. Nat. Hist. (8) **20**: 328–350.

Schlaikjer, Erich M.

1935. Contributions to the stratigraphy and paleontology of the Goshen Hole area, Wyoming. IV. New vertebrates and the stratigraphy of the Oligocene and early Miocene. Bull. Mus. Comp. Zool. Harvard Coll. **76**: 97–189.

Scott, W. B., & Jepsen, G. L.

1936. The mammalian fauna of the White River Oligocene—Part I. Insectivora and Carnivora. Trans. Amer. Phil. Soc., n. s. **28**: 1–153.

Severtzow, M. N.

1857 1858. Notice sur la classification multisériale des carnivores, specialement des felidés, et les études de zoologie générale que s'y rattachent. Rev. Mag. Zool. (2) **9**: 387–391, 433–439, (2) **10**: 3-8, 145-150, 192–199, 241-246, 385–393.

Simpson, George Gaylord

1937a. Super-specific variation in nature and in classification from the view-point of paleontology. Amer. Nat. **71**: 236–267.

1937b. Notes on the Clark Fork, Upper Paleocene, fauna. Amer. Mus. Novitates, No. **954**.

1941. Range as a zoological character. Amer. Jour. Sci. **239**: 785–804.

Stirton, R. A.

1940. Phylogeny of North American Equidae. Univ. California Pub., Bull. Dept. Geol. Sci. **25**: 165–198.

Waagen, W.

1869. Die Formenreihe des Ammonites subradiatus. Benecke Geog.-pal. Beitr. **2**: 179–257.

CRITERIA FOR SPECIES AND THEIR SUB-DIVISIONS FROM THE POINT OF VIEW OF GENETICS

By

W. Frank Blair

University of Michigan, Ann Arbor, Michigan

INTRODUCTION

The problem of the evolution of species has been much clarified in recent years. Progress has been possible principally because the data from the separate disciplines of genetics, ecology, biogeography, and morphology all have been brought to bear on the problem. The present discussion pertains to the evolution and classification of the vertebrates. The evidence to be presented is mostly from the field of mammalogy, partly because of the author's specialization in that field and partly because of recent progress toward an understanding of the evolution of mammalian species.

Most modern geneticists, with the notable exception of Goldschmidt (1940), agree that species develop through isolation and the gradual accumulation of minor mutations in the isolated stocks These mutations, of course, may affect the physiology of the stocks as well as their physical characters. This is speciation through microevolution. The opposing view of Goldschmidt, that species arise by macroevolution—that is, through sudden, major, or systemic mutations--cannot be discussed here for want of time. Suffice it to say, however, that most geneticists are convinced that speciation occurs through microevolution and that the evidence to be presented here supports this view.

RACIATION

Most species of vertebrates, if they are at all widespread, are divisible into numerous geographical races. Many of the more obvious of these races have been named; they constitute the subspecies of vertebrate systematists. Each geographical race is in contact with one or more other races of the same species. These races are fertile each with the others (Dice, 1933), and along the contacts they form intergrading populations. Where contact is prevented by ecological barriers, the races concerned are connected by intergradation through chains of races.

Consequently, a mutation that arises in any geographical race theoretically could be dispersed ultimately to all of the other races.

The differences between geographical races are hereditary, and the genetic differences between races are no different than those between individuals. This first was proved by Sumner (1932) and later confirmed by Dice (1937 and other papers). There is some evidence (author's unpublished data) that the tremendous raciation in at least one species (*Peromyscus maniculatus*) involves but a relatively small number of genes. How then do we account for the existence of numerous geographical races if the stock of genes is relatively small and there is the opportunity for transfer of these genes to all parts of the interbreeding population? The logical explanation is that geographical races indicate ecological trends (Dice and Blossom, 1937). From the stock of genes available in the entire population of the species or incipient species, selection has, in any given environment, weeded out the genes that are non-adaptive for that environment and has conserved those that are adaptive. Selection pressure will act to maintain the most successful genetic combination in spite of gene mutation and the spread of genes from other races of the same species.

As the environment varies geographically so will the most successful gene complexes vary geographically. The geographical races that have been named represent, for the most part, major ecological trends. Within the area of one of these major trends many minor variations in environment occur, and, likewise, many local variations in gene complexes (see Dice, 1940a). Thus, a geographical race usually is but a part of the species population that is adapted to a particular environment, and many locally adaptive variations may occur within its limits. Geographical races also may be produced as a result of the random drifting apart of small, partially isolated colonies (Wright, 1931).

Raciation and speciation are distinctly different evolutionary processes. Geographical races are not necessarily incipient species (Dice and Blossom, 1937; Goldschmidt, 1940; Wright, 1940). The process of speciation is initiated by the isolation of a part of the previously interbreeding population. The part of the population becoming isolated might conceivably comprise a geographical race, but it might comprise, instead, only a part of a geographical race or several races. If the split did occur along racial lines there would be an illusion of rapid divergence because of the initially different gene complexes of the two populations. However, there is no reason whatsoever for believing that under such conditions isolating mechanisms would be developed any more rapidly or that

infertility would appear any sooner than if the split had occurred independently of racial lines.

SPECIATION

Speciation occurs as the result of: (1) isolation of one or more parts of a previously interbreeding population, (2) morphological differentiation as the result of differential mutation and selection pressure, and (3) the development of mutual infertility through genic or chromosomal changes.

The process of species differentiation is reversible up to the point at which the diverging populations become so different genetically that the interchange of genes is no longer possible. Geographical barriers may be overcome by the spreading out of one or the other of previously isolated populations until the two populations merge. Ecological isolation may break down. A. P. Blair (1941) believes that man-made ecological changes account in part for the present hybridization in nature of certain toad populations. These previously had so diverged morphologically and ecologically as to be ranked as taxonomic species. He suggests that this may be a case of fusion and disintegration of species. Psychological barriers are only relatively effective, and they may be overcome under certain conditions of population pressure. Hubbs and Hubbs (1932) believes that hybridization of sunfishes occurs as the result of intensive population pressure and limited spawning grounds in certain types of pools. Miller (1941) found numerous instances of hybridization of juncos due to the partial failure of one or more of these isolating mechanisms.

To maintain a completely dynamic point of view, we must remember that the process of speciation can be reversed so long as isolation is maintained only by geographical, ecological, or psychological barriers. One of these isolating mechanisms or a combination of several may operate at any one moment to effectively bar the interchange of genes between two incipient species, but there is no certainty that future events will not break down the barrier. It is only when the interchange of genetic material between the populations is made impossible by infertility that an irrevocable step in speciation is taken.

The conventional taxonomic system obviously does not distinguish between the differentiating populations that have, and those that have not, passed the point of irreversibility. A dynamic system of classification is needed, therefore, to show the evolutionary relationships of these natural populations. We propose to utilize such a system here. In this system, the criteria for discriminating between the different classificatory units are genetic. Strictly speaking, these criteria concern the

ability of different evolutionary units to exchange germinal materials with other units. Applying these criteria, we have two categories of species populations: (1) those that are isolated from all other populations by reason of sterility and (2) those that are isolated at any given moment by mechanisms short of intersterility.

Incipient species—that is, diverging populations that have not yet attained intersterility—we propose to call just what they are, incipient species. As we define it, an incipient species is *a natural population that is at least partially fertile with some other population but is inhibited from breeding with it by some isolating mechanism or mechanisms.* For populations that have reached intersterility, a satisfactory term already exists in botanical literature. The *cenospecies* as proposed by Turesson (1922) and used by Gregor, Davey, and Lang (1936), Clausen, Keck, and Hiesey (1939), and others corresponds to this category in our system of classification. The cenospecies, as redefined here to apply to animals as well as plants, is *a natural population that is infertile (can produce only sterile hybrids or none at all) with every other population.*

In setting up a dynamic system of classification that is complementary to the orthodox system instead of attempting to change the latter system to fit the experimental data, we have followed the example of the experimental botanists. These workers probably followed a wise course in not attempting to revise the orthodox system during the early stages of experimental taxonomic research. However, such a course merely postpones the day when a major revision of the conventional taxonomic system must be made to utilize the criteria furnished by the experimental method. It seems to the present author that the problem of revision is one for the taxonomist.

The incipient species of our classification corresponds in many, but not all, cases to the taxonomic species. The cenospecies, likewise, corresponds in many, but not all, cases to the species-group, an informal category, of systematic mammalogy. However, some taxonomic species are conterminous with the cenospecies. Until such time as the conventional system of classification is revised to use the criteria furnished by the experimental method, the term species must be restricted to museum species, that is those species units based on morphological and geographical criteria alone. It should be kept in mind that these museum species do not necessarily represent evolutionary units. The terms incipient species and cenospecies are to be used only when the evolutionary relationships of the populations concerned have been established by experiment.

A nomenclatorial problem arises here, because no formal category of the order of the cenospecies is recognized in orthodox systematics. In

the following discussion the name of each cenospecies is made to coincide with the first formally recognized species name within its limits. This usually corresponds to the name of the species-group as used informally in mammalian taxonomy.

Speciation in *Peromyscus*

The species-group of mammalian taxonomists has been shown by experiment to fit our definition of a cenospecies in at least some cases. Dice (1933) has shown that in all crosses attempted between subgroups of a species-group at least some fertile offspring were produced, but no offspring were obtained in attempted crosses between species-groups.

Speciation in *Peromyscus* has been studied more intensively than in any other group of mammals, due primarily to the early work of Sumner (1932, for list of publications) and the work of Dice (1940a, and numerous other papers) and his collaborators. Nine apparent cenospecies of *Peromyscus* occur in North America north of Mexico. Four of these cenospecies, *californicus*, *crinitus*, *nuttalli*, and *floridanus*, have no subgroups of significance in speciation, although most have undergone some raciation. Two other cenospecies, *boylii* and *eremicus*, are each split into two apparently separate breeding arrays, each with its own geographic races. In the first of these cases, the two arrays, *boylii* and *pectoralis*, occur together in the same regions and in the same ecological communities However, the fertility relationships in this case, and in the case of the two subgroups of *eremicus*, *eremicus* and *merriami*, are as yet obscure. The three remaining cenospecies of *Peromyscus* have provided most of our evidence about the course of speciation in this genus.

The cenospecies *leucopus* is split into two separate breeding arrays, which have been named, respectively, *leucopus* and *gossypinus*. The range of *leucopus* extends from southern Mexico north to Montana and east to Nova Scotia, but it does not extend into southern Alabama, Georgia, South Carolina, nor into any part of Florida. On the other hand, *gossypinus* ranges from Florida west to eastern Oklahoma and Texas, north to Tennessee, and east to southern Virginia. The ranges of the two populations overlap in a broad strip extending from eastern Texas and Oklahoma to Virginia. The *leucopus* and *gossypinus* cross freely in the laboratory and produce fertile offspring (Dice, 1937a). However, Dice (1940b) found no evidence of hybridization between the two arrays where they occurred together in the Dismal Swamp region. Osgood (1909) found in museum materials no evidence of hybridization of the two, and consequently treated them as taxonomically distinct species.

The *leucopus* and *gossypinus* arrays constitute two separate incipient species in our system of classification. The two arrays combined comprise a cenospecies, which is recognized to be split into the two diverging populations. The separation of the two populations probably occurred in the not distant geological past. The *gossypinus* population could possibly have become separated from the parent *leucopus-gossypinus* population during one of the Pleistocene inter-glacial periods, when, due to a raised sea-land, much of the peninsula of Florida existed as a large island (see Cooke, 1939). When, with a lowering of the sea level, this island again became connected with the mainland the opportunity arose for this population to spread out over the southeastern coastal plain. Today, this spreading has reached the point where the two populations overlap broadly. During the course of the separation, however, the two populations have diverged morphologically to the extent that they differ in the size of certain parts of the body. The morphological differences between the two arrays, however, are no greater than those that exist between some races of the *leucopus* array. The most important divergence between the two populations, though, seems to have been a psychological one. This acts as an isolating mechanism to prevent interbreeding now that the home ranges of the two arrays overlap. So long as this isolating mechanism effectively prevents interchange of genes between the two populations the two are free to drift apart through differential mutation and selection. No infertility yet exists between the two, however, so the process of differentiation still is reversible. The psychological chasm between them possibly may yet be bridged under some conditions of population pressure. In our classification, the populations of *leucopus* and *gossypinus*, therefore, are to be considered incipient species.

The cenospecies *maniculatus* is broken into several apparently discrete breeding arrays. The genetic relationships of only two of these arrays, *maniculatus* and *polionotus* have been investigated. The *maniculatus* population ranges over most of North America from southern Mexico to Alaska and Labrador, but like the *leucopus* array of the cenospecies *leucopus* it does not range into the southeastern corner of the United States. It is replaced there by a geographically isolated, morphologically differentiated array, *polionotus*. The separation of these two arrays within the cenospecies *maniculatus* probably was brought about by the same event that split the cenospecies *leucopus* into two separate populations. Actually, the *polionotus* array comprises not one geographically isolated breeding population but three or more, for populations of these mice occur on at least two islands. The island populations are as effectively isolated

from the mainland population as that population is from the *maniculatus* array. The island populations, the subspecies *leucocephalus* and *plasma* of systematic mammalogy, must be considered incipient species under our system, just as the mainland population must be considered such.

The fertility relations of these populations, in so far as we know them, are extremely interesting. The mainland population of *polionotus* crosses with the *maniculatus* population in the laboratory and produces fertile offspring (Watson, 1942). Furthermore, the *leucocephalus* population from Santa Rosa Island, Florida, crosses readily with the mainland population of *polionotus* and produces fertile offspring (Sumner, 1930). It would seem, thus far, that no infertility has appeared between any of these incipient species. However, in crosses in which a laboratory stock combining the germinal materials of the Santa Rosa Island and mainland populations was mated with representatives of the *maniculatus* array some unexpected results were obtained (author's unpublished data). Only a few of the matings produced offspring. The F_1 animals were only partially viable, and many died shortly after birth. The sex ratio was unbalanced significantly in favor of females. The F_1 females were fertile, but both fertile and sterile F_1 males were produced. The sterility of some F_1 males appears to be due to gross disturbances in spermatogenesis (this is being investigated by Moree, unpublished). The logical explanation is that the *leucocephalus* population contributed the partial infertility with *maniculatus*, since the mainland *polionotus* population has proved fertile in crosses with *maniculatus*. This is being further investigated by crossing pure *leucocephalus* mice with *maniculatus*.

If our assumption is correct that the partial infertility in the above-mentioned cross came from the *leucocephalus* population, then that population has diverged farther from the *maniculatus* array than has the mainland *polionotus* population. Physically, the differences exhibited by the *polionotus*, *leucocephalus* and *maniculatus* populations are no greater than the differences between some of their geographic races.

The *leucocephalus* population differs genetically from *maniculatus* in at least one important adaptive character. A single unit factor for the dorsal extension of ventral white (white cheek, in author's unpublished data) is dominant over the "normal" condition found in *maniculatus*. There also is, in *leucocephalus*, a series of modifiers that act to extend the white progressively farther and farther onto the dorsal surface (see Sumner, 1930). All *leucocephalus* that have been examined are, phenotypically at least, white cheek, and all representatives of the *maniculatus* array are "normal." In the mainland *polionotus* population, littoral races are genetically white cheeked, while interior races are "normal"

If, in the course of future events, *leucocephalus* should develop complete infertility with *maniculatus* and the mainland *polionotus* population should become extirpated, then the great difference between the white cheeked *leucocephalus* and the "normal" *maniculatus* would constitute one of the so-called "bridgeless gaps" on the basis of which Goldschmidt (1940) attempts to discredit speciation through microevolution. The gap would be bridgeless, of course, only at that hypothetical future date and only after the connecting links had disappeared.

The cenospecies *truei* affords the last example of speciation in progress that we will discuss. In the southwestern United States this cenospecies is split into two distinct breeding arrays, the taxonomic species *truei* and *nasutus*, each with several geographic races. The geographic ranges of these two arrays overlap broadly, and in many places the two occur together in the same ecological associations (Dice, 1942). The two arrays can be separated easily on the basis of morphological characters, of which the most distinctive are the size of the external ear, size of the auditory bullae, and relative length of the tail. Representatives of the two arrays can be crossed in the laboratory. The F_1 females are fertile, but the F_1 males all are sterile (Dice, 1937b). No hybrids between the two have been found in nature. It seems evident, therefore, as Dice (1942) has pointed out that some psychological barrier must prevent breeding between the two populations.

The *truei* and *nasutus* populations represent a closer approach to mutual infertility, and consequently to the irreversible stage in speciation, than is evident in the other cases of incipient speciation that we have discussed. The lowest stage of speciation in our material has been reached by the *leucopus* and *gossypinus* arrays, which, apparently while geographically isolated, have diverged morphologically and have developed a psychological barrier sufficient to prevent interbreeding now that the populations are in contact. The divergence between the *maniculatus* and *polionotus* populations is of approximately the same order, and so is that between *polionotus* and *leucocephalus*. The differentiation between the *leucocephalus* and *maniculatus* represents a further step in speciation, for partial infertility has developed. The *truei* and *nasutus* populations represent a still further step, for in this case all hybrid males are sterile. When the hybrid females, too, become sterile the divergence of these populations one from the other will have passed the point from which there is no turning back. However, until that point is reached there always is the possibility that the psychological isolating mechanisms may break down and thus permit the fusion of the two

populations. Therefore, *truei* and *nasutus* still are but incipient species under our dynamic system of classification.

The incipient species of *Peromyscus* exhibit several noteworthy evolutionary trends. In the two cenospecies *leucopus* and *maniculatus*, geographic isolation appears to have been the first step in speciation. The evidence admittedly is circumstantial, but it seems clearly indicated that within these cenospecies "physiological isolating mechanisms" (see Dobzhansky, 1941: 257) were developed only after geographic isolation had taken place. If these cases are representative, then geographic isolation is of paramount importance in speciation. A similar view of the importance of geographical isolation is held by Miller (1941) in respect to speciation in juncos. The principal morphological differences in the diverging populations, both in *leucopus* and *maniculatus*, are in size alone and apparently are non-adaptive. The morphological differences between the diverging arrays of the cenospecies *truei*, which represent later stages of speciation, are differences of proportion as well as of size, and probably are in part adaptive. There is at least a suggestion, therefore, that the first divergence in isolated populations may be due to chance drifting apart in non-adaptive characters.

SUMMARY

Geographical races usually indicate ecological trends; hence they are not necessarily incipient species. Speciation comes about through the isolation, differentiation, and ultimate intersterility of parts of a previously interbreeding population. The process of speciation is reversible up to the point at which the exchange of genes becomes no longer possible because of intersterility.

In a system of classification that makes use of genetic and ecological criteria, the cenospecies is defined as a population that is infertile with every other population. Any isolated population that has not evolved far enough to be infertile with related populations is regarded as an incipient species. Such a system shows relationships and evolutionary trends better than does the conventional taxonomic method.

LITERATURE CITED

Blair, A. P.

1941. Variation, isolation mechanisms, and hybridization in toads. Genetics **26**: 398–417, 6 figs.

Clausen, J., Keck, D. D., & Hiesey, W. M.

1939. The concept of species based on experiment. Amer Jour. Botany **26**: 103 106.

Cooke, C. W.

1939. Scenery of Florida interpreted by a geologist. Fla. Geol Bull. **17**: 1 118, 58 figs.

Dice, L. R.

1932. Variation in a geographic race of the deer-mouse, *Peromyscus maniculatus bairdii*. Occas. Pap. Univ. Mich. Mus. Zool. **239**: 1–26, 1 fig.

1933. Fertility relationships between some of the species and subspecies of mice in the genus *Peromyscus*. Jour. Mammalogy **14**: 298–305.

1937a. Fertility relations in the *Peromyscus leucopus* group of mice. Contr. Lab. Vert. Genetics **4**: 1–3.

1937b. Partial infertility between two members of the *Peromyscus truei* group of mice. Contr. Lab. Vert. Genetics **5**: 1–4.

1940a. Ecologic and genetic variability within species of *Peromyscus*. Amer. Nat. **74**: 212–221.

1940b. Relationships between the wood-mouse and cotton-mouse in eastern Virginia. Jour. Mammalogy **21**: 14–23, 1 fig.

1942. Ecological distribution of *Peromyscus* and *Neotoma* in parts of southern New Mexico. Ecology **23**: 199–208, 1 fig.

Dice, L. R., & Blossom, P. M.

1937. Studies of [illegible]malian ecology in southwestern North America with special att[illegible]o the colors of desert mammals. Carnegie Inst. Wash. Pub[illegible]+1–129, 8 figs., 8 pls.

Dobzhansky, T.

1941. Genetics and the origin of species. New York: xviii+446 pp., 24 figs.

Goldschmidt, R.

1940. The material basis of evolution. New Haven: xi+436 pp., 83 text figs.

Gregor, J. W., Davey, V. M., & Lang, J. M. S.

1936. Experimental taxonomy. I. Experimental garden technique in relation to the recognition of the small taxonomic units. New Phytologist **35**: 323–350, 3 figs.

Hubbs, C. L., & Hubbs, Laura C.

1932. Experimental verification of natural hybridization between distinct genera of sunfishes. Papers Mich. Acad. Sci., Arts, Letters **15**: 427–437.

Miller, A. H.

1941. Speciation in the avian genus *Junco*. Univ. Calif. Publ. Zool. **44**: 173–434, 33 figs.

Osgood, W. H.

1909. Revision of the mice of the American genus *Peromyscus*. North Amer. Fauna **28**: 285 pp., 12 figs., 8 pls.

Sumner, F. B.

1930. Genetic and distributional studies of three subspecies of *Peromyscus*. Jour. Genetics **23**: 275–376, 27 figs., 11 pls.

1932. Genetic, distributional, and evolutionary studies of the subspecies of deer mice (*Peromyscus*). Bib. Genetica **9**: 1–106, 24 figs.

Turesson, G.

1922. The genotypical response of the plant species to the habitat. Hereditas **3**: 341–347.

Watson, M. L.

1942. Hybridization experiments between *Peromyscus polionotus* and *Peromyscus maniculatus*. Jour. Mammalogy **23**: 315–316.

Wright, S.

1931. Evolution in mendelian populations. Genetics **16**: 97–159, 21 figs.

1941. The material basis of evolution (review). Sci. Monthly **53**: 165–170.

Annals of the New York Academy of Sciences
Volume XLIV, Art. 3. Pages 189–262
September 30, 1943

PARASITIC DISEASES AND AMERICAN PARTICIPATION IN THE WAR*

By

Horace W. Stunkard, Lowell T. Coggeshall, Thomas T. Mackie, Robert Matheson, and Norman R. Stoll

CONTENTS

*This series of papers is the result of a conference on Parasitic Diseases and American Participation in the War held by the Section of Biology of The New York Academy of Sciences, March 18, 1943.
Publication made possible through a grant from the income of the Conference Publication Revolving Fund.

INTRODUCTION TO THE CONFERENCE ON PARASITIC DISEASES

By Horace W. Stunkard

New York University, New York, N. Y.

For some time the officers of the Section of Biology have contemplated a conference dealing with the biology of animal parasites, and although this subject is of primary and fundamental importance, it was felt that consideration of more immediate and practical aspects of parasitology might be of more value at this time. The study of animal parasites can no longer be regarded merely as an academic subject; the diseases produced by these organisms present one of the momentous and pressing problems of today and tomorrow. The dispersal of American military and naval personnel to all parts of the world, especially to tropical and subtropical areas, and their operation under field conditions, has created an increasingly grave problem not only to the officers responsible for the health and efficiency of our troops but to the civilian population as well. The timeliness of this conference is apparent to members of this group and requires no defense or justification. It is indicated by a widespread interest in tropical medicine of which parasitic diseases constitute the major component. This interest is manifested in many ways and was expressed in the symposium arranged for December 29, 1942, jointly by the American Association for the Advancement of Science, The American Society of Parasitologists, The National Malaria Committee, The American Society of Tropical Medicine and The New York Society of Tropical Medicine. Although the symposium was cancelled at the request of the Office of Defense Transportation, certain of the papers were presented under other auspices and the Theobald Smith lecture of the New York Society of Tropical Medicine by Dr. Ruiz Castañeda was attended by many of the present audience. The urgency for a restatement of present knowledge and for consideration of immediate research projects in the realm of parasitic diseases is so obvious that the officers of the Section of Biology are to be congratulated for arranging this conference.

It is my assignment, and I find it a very agreeable one, to welcome you here today. On behalf of The New York Academy of Sciences, I wish especially to express gratitude and appreciation to the men who have so generously contributed their time to the preparation of the formal papers for this program. All of them are eminent in their

respective fields and all are busy with added duties in connection with the war effort. It is the belief of the Committee that uninhibited and frank discussion will prove stimulating to members of the conference and that it will suggest new avenues of approach for the investigation of current problems. An invitation to attend was extended to the members of the New York Society of Tropical Medicine and I am happy to greet so many members of our local Society.

It may be appropriate and instructive to compare the training and experience in the treatment, control and prevention of parasitic diseases afforded members of the medical profession in the United States and in those countries with which we are associated and those against which we are opposed in the war. The medical departments of the armies of Germany, Italy and Japan have long been preparing for the conflict and extensive studies had been carried on before the outbreak of hostilities. The staff of the British army, likewise, was familiar with the problems and prepared to meet them. Indeed, through their colonial interests, the British, German, French, Dutch and Belgians have long recognized the importance of parasitic diseases. The trail was blazed by medical missionaries, a small but unselfish and devoted band of men who, without particular preliminary training, learned about parasitic diseases at first hand and by empirical methods. The development of commercial relations between the homeland and the colonies, the migration of Europeans to the colonies and their subsequent return, the more or less constant intercourse through shipping, the necessity of maintaining healthful conditions for Europeans and natives in the colonies, all have combined to bring the existence, prevalence and importance of parasitic diseases to the attention of the medical profession in these countries. As a consequence, the medical schools have developed strong departments of parasitology either as integral parts of the institutions or as associated institutes. The schools of London and Liverpool were founded in 1899 and that of Hamburg the following year. The institutes of colonial medicine in the universities of Paris, Antwerp and Amsterdam have long provided special training in this field. It has been customary for young physicians, trained in European schools, to spend some time in the colonies and many of them have devoted their lives to this branch of medicine. The names of the more noteworthy are so well known that it would be redundant to enumerate them here. To further the instruction in parasitic diseases, these medical schools have built up museums and teaching collections, and members of their staffs have been sent to different parts of the world to assemble teaching material. Not only have they been engaged in teaching, but each of the institutes named—

and more particularly the Molteno Institute for Research in Parasitology, at the University of Cambridge—has maintained an active research program.

In contrast to the situation in Europe, the entry of the United States into the present war found the medical profession unprepared to cope with the problems which are arising as increasing numbers of American troops are sent into areas where tropical diseases are endemic. Possibly the majority of our troops in this war will serve in regions where the native population is heavily infected, where the probability of infection is almost a certainty, and where, under field conditions, preventive and sanitary measures must be inadequate at best. I think it would be a conservative estimate to say that some million or more Americans may acquire parasitic diseases, and the loss of efficiency and striking power on the part of the armed forces will be very greatly reduced. Indeed, it has been reported that our defeat at Bataan was due as much to malaria and other diseases as to the military factors concerned. The more or less constant repatriation of troops, incapacitated in the field, will mean a continuous importation of parasites that constitutes a menace to the health of the United States. Many of these diseases will require weeks or even months for the development of symptoms and the presence of the infections may not be recognized for some time after the troops are home. In the case of those diseases which are spread directly, only climatic factors and sanitary measures will limit their establishment, once they are introduced. In the case of those organisms which require an intermediate host for the completion of the life cycle, it is possible that some of them may find in the United States suitable vectors other than the original one. It is well known that, in the course of time, parasites do acquire new intermediate as well as definitive hosts. The sheep liver fluke is probably a classic example. Carried to all parts of the world by human migrations with their domestic stock, the parasite has found suitable new intermediate hosts in the many countries into which it has been introduced. In this country there are established schistosome flukes of birds, whose larvae cause swimmer's itch, and I think it not unlikely that there are species of snails which may become intermediate hosts for human schistosomes. Individuals of every species manifest variations and some variance of the parasite or of the host may permit completion of the cycle and establish the parasite. Of those parasites which require insect or other arthropod vectors, the possibility of finding a suitable intermediate host is great and there may be as yet unrecognized species capable of transmitting the parasite. Indeed, in conversation, Dr. L. E. Rozeboom stated recently that he had found an

undescribed species of *Phlebotomus* in southern United States, which suggests the existence of possible infecting agents concerning whose presence we are not even aware. Not only is the situation serious from the point of view of men actually in service, but war conditions have in the past and presumably may now lead to outbursts of latent infections in the general population. More details concerning these subjects will undoubtedly be presented by the speakers on today's program.

The distressing situation with regard to tropical medicine and parasitic diseases, as it applies to our armed forces, is in large part due to the fact that in the past there has been little necessity for American physicians to familiarize themselves with parasitic diseases and that American medical schools and colleges have not given adequate instruction in tropical medicine. It is no secret that our medical instruction on the subject of parasitic diseases has been of the most desultory character. With notably few exceptions, the diseases caused by animal parasites have been considered incidentally in courses devoted primarily to other subjects, such as bacteriology, pathology or public health. Independent departments of parasitology, staffed by experts, have never been developed in American medical schools. European workers have often expressed amazement that in the United States there is no institution devoted primarily to instruction and research in parasitic diseases. Nuttall, Brumpt and Fülleborn have voiced the opinion that since New York is the principal shipping and commercial center of the United States, it should provide adequate clinical material and financial support for such an institution. In time of war it serves as one of the chief ports of embarkation and debarkation and the need for a diagnostic, therapeutic and research laboratory becomes acute. The urgency will increase during the war and postwar period and provision should be made at once for the organization and establishment of a center for the investigation of parasitic diseases.

In these introductory remarks, I have tried to sketch the general outlines of the situation, to enumerate a few of the more outstanding features, and to suggest at least one step that should be taken to meet current problems.

CURRENT AND POSTWAR PROBLEMS ASSOCIATED WITH THE HUMAN PROTOZOAN DISEASES

By L. T. Coggeshall
University of Michigan, Ann Arbor, Michigan

INTRODUCTION

The present war is providing unparalleled opportunities for the entrance into this country of pathogenic agents and disease vectors that have played little or no part in our past medical history. Chief among the diseases most likely to exert their influence over us is the protozoan group and their near relatives which thrive in tropical and subtropical areas because of favorable climatic conditions and lack of adequate control. In the past, some have had their distribution limited to fairly restricted areas, largely because of climatic and physical barriers which affect host and parasite alike. However, the majority have a very widespread distribution and would seem to need only slight aid to extend their boundaries into unaffected but receptive territory. A major portion of our troops on foreign soil is now in highly disease-ridden areas. In some places they are acquiring infections from the native reservoirs at an alarming rate. By the rehabilitation of sick troops these recently contracted diseases are being transposed to the United States in large numbers. Whether the infections succeed in establishing themselves in their new environment is yet to be learned but we cannot assume that they will disappear spontaneously because our soil is unfriendly. Therefore it must be agreed that all are potentially dangerous and unless energetic measures are instituted to curb their activity we may face serious consequences. In this discussion I should like to present evidence for considering these foreign diseases and their vectors as immediate hazards to present and future health.

BIOLOGICAL BEHAVIOR OF PROTOZOAN DISEASES

To understand why the Protozoa in particular and their relatives, the filariae and schistosomes, are of special importance one needs only to review their biological characteristics. If I were asked what features distinguish them as a group I would reply that in general it would be their tendency to produce chronic infections and their inability to leave more than a transitory immunity in their respective hosts. With the excep-

tion of the amoebae and the intestinal flagellates, the pathogenic agents under discussion have complicated life histories in man and their intermediate insect hosts. The malarial plasmodia, trypanosomes, leishmaniae, filariae, and schistosomes possess strict requirements for survival which would seem to make them liable for extinction. On the other hand, they have compensated for this handicap by their capacity to exist almost indefinitely in man. It is to their advantage not to destroy their host but to establish a permanent host-parasite relationship. For example, both the amoebae and the intestinal flagellates persist almost indefinitely in the intestinal tract of man. The malarial plasmodia go through no less than twelve distinct morphological stages in man and mosquito. In the latter the infection lasts only a few weeks but in the human host it runs a very chronic course with a tendency to relapse repeatedly. Actually there are cases on record where malarial infections have lasted a score or more years. For example, there was a Greek in Denver who remembered having had a malarial infection in Greece shortly before he became an immigrant to this country at the age of ten. After residing 37 years in Denver, where malarial transmission does not occur, he gave his blood for transfusion purposes and the recipient came down with quartan malaria. Upon examination of the donor's blood circulating parasites were found, yet he had never had any symptoms referable to the disease in the interim between the initial infection and the transfusion episode. There are many other similarly authenticated reports, usually involving the quartan parasite, the most chronic of the malarial infections. Vivax and falciparum malaria are of shorter duration, but they do not die out when the acute infection subsides. Any one of the returning infected soldiers, the majority of whom will have the dreaded falciparum malaria because they are largely in areas where that species is predominant, conceivably can serve as a focus for many months or even years The vector for malaria in the United States, *Anopheles quadrimaculatus*, is abundant as far north as the Canadian border. It is capable of transmitting all of the human plasmodia and will not need to depend upon a single exposure to become infected, but will have repeated opportunities to feed upon chronic or relapsing cases.

Trypanosomiasis is one of the dreaded tropical diseases and thus far has been confined largely to the tropics. The African form, sleeping sickness, is limited to the distribution of its vectors, the *Glossina* or tsetse flies. Both the African and Western Hemisphere varieties produce long-standing infections. In the former the terminal stages usually are not reached until the third year. In the latter, Chagas, for whom the disease was named, believed that if an individual did not die in the

initial attack he would develop a chronic infection that would last indefinitely.

Leishmaniasis, both the visceral and cutaneous variety, known as kala azar and oriental sore respectively, is a chronic infection. Its duration is especially prolonged. A variety of oriental sore, known as espundia, is present in parts of the Western Hemisphere, and has the same chronic characteristic.

The filarial organism, *Wuchereria bancrofti*, that causes elephantiasis, rarely subsides of its own accord but establishes a relationship with the human host that persists for many years. Death intervenes only when secondary infections occur.

Schistosomiasis, so prevalent in Africa, the Middle and Far East, and South America, is another example of a very chronic disease. Any infected individual is capable of releasing thousands of ova from the intestinal or urinary tract for years.

Thus we can readily see that this group of diseases possesses as one of its special behaviors the tendency to produce infections of long duration. Whether in their evolutionary development it has been necessary for them to acquire this characteristic in order to survive is not known. However, if introduced into new areas we do know by this same feature that they can serve as sources of infection over long periods and thus enhance their chances for spreading.

The nature of the immunological reactions produced in man as the result of infection with the above-mentioned diseases furnishes some information on the probability of their danger to us. As a group they probably induce no more than a transitory immunity. For example, when amoebic infections disappear spontaneously or are cured by specific therapy there is no residual immunity and there can be subsequent attacks as severe as the initial one. The same is true of malarial infections as they confer no lasting immunity upon their hosts after complete recovery. An attack of any one of the three human plasmodia offers no protection against the other two. As a matter of fact there is no cross immunity between strains of the organism within the same species. James[1] was the first to show that after the acute attack of malaria had subsided the individual was highly immune to the homologous strain of vivax malaria, yet as susceptible as any normal when inoculated with the same species imported from a distant area. Less is known about the immune responses to the human trypanosomes or leishmaniae but there is little or no evidence of a permanent efficient

[1] James, S. P. Some general results of a study of induced malaria in England. Trans. Roy. Soc. Trop. Med. and Hyg. 24: 477. 1931.

immunity response and it is possible to produce experimentally repeated attacks in some of the lower animal hosts. It seems significant that as yet there are no effective vaccines for any of the protozoan infections. If one is permitted to make a generalization on this point it would be that diseases associated with the minimum duration of immunity offer the least hope for artificial immunization. If this group of diseases conferred permanent immunity after the initial attack then their hosts would not serve as an indefinite source of danger.

ACQUISITION OF INFECTION IN TROPICAL AREAS

Another important factor is the acquisition of the various infections in tropical zones. As a country without colonies we have had relatively few contacts with the tropical world, with the exception of Panama, the Philippines and a few Caribbean Islands. Even a considerable proportion of our maritime commerce has been conducted by foreign crews and ships. With the advent of the war it suddenly became necessary for us to transport huge numbers of troops overseas and a major part of them is now in tropical areas. For the most part they are in places that enjoy the dubious reputation of being the foremost disease centers of the world—Africa, the Middle East, India, China and the Southwest Pacific. If we ever approach the proposed figure of 11,000,000 men in the armed services, approximately one-half of our male adult population, and if the same proportions are to serve in the tropics as now, then we can gain some appreciation of the probable consequences, if only a small percentage acquires one or more of the tropical infections. We need not speculate on the probability of these troops becoming infected, because it is already a fact.

Malaria and dysentery are and will undoubtedly continue to be the chief offenders. In the latter group the amoebae are important because of their prevalence and widespread distribution. Amoebae occur endemically throughout all the tropical countries and in most areas cause the prevailing form of dysentery. The incidence of infection found by survey in various native populations usually varies between 10 and 25 per cent. On the Gold Coast in West Africa it was found by routine monthly stool examinations that approximately 12 per cent of the native Africans employed as food handlers along an American air route were infected with *Endamoeba histolytica*. This percentage was increased to 20 when purgation was used to obtain liquid stools. In approximately 1000 airline personnel, only 16 acute cases developed during a year's operation. Surveys throughout the United States show that 5 to 10 per cent are carriers of *E. histolytica*, so the dangers of intro-

ducing new strains should not cause much apprehension unless they are much more virulent. On the other hand possibilities of serious outbreaks are not remote as witnessed by the severe Chicago epidemic which embraced at least 1400 severe cases and 52 deaths.

Malaria is the predominant disease in the tropics. In Liberia, for example, repeated surveys were made among the natives in villages adjacent to our troops. The average number of children showing circulating parasites was near 100 per cent, while in adults the figure averaged 70 per cent. These rates are high but not excessive for Africa. Our troops are highly susceptible to this infection. In my own experience with the personnel of a large airline, during one month over 40 per cent were incapacitated within eight weeks of their arrival in a West African country.[2] In combat areas where control is very difficult, malaria reaches epidemic proportions. At the fall of Bataan it was estimated that 85 per cent of each regiment had acute malaria. In some areas of the Southwest Pacific the incidence is almost as high. Recent arrivals of United States troops from these areas show that a high percentage of the medical cases are suffering from malaria or have recently recovered from its effects.

Leishmaniasis is not just a term for an exotic disease that occurs sporadically. Its incidence is very high especially in Africa and the Middle and Far East. An average of 60,000 cases are treated annually in Assam province. This represents only a fraction of the actual number of cases. In 1937, in Bengal, 137,000 individuals presented themselves for treatment. In Africa the chief centers are in Morocco Algeria, Tunisia, Tripolitania, and Egypt. These names are significant to all of you. Recently the disease has been reported in the Western Hemisphere especially in South America.

Trypanosomiasis occurs in the proximity of our troops in several parts of Africa. Although it is not so prevalent as the other protozoan diseases, there is an appreciable number of cases. For example, in Liberia the rates are about one in 500 in the coastal natives. White persons are susceptible, but less so than the negroes, apparently because of the preferential biting habits and opportunities for bites of the tsetse fly.

Schistosomiasis, although not a protozoan infection, presents similar problems and should be mentioned. It has a widespread distribution in the Mediterranean countries, Middle and South Africa, and the Near and Far East. During the early slave-trading days, schisto-

[2] Coggeshall, L. T., et al. Experiences in the development of a medical service for airline operations in Africa. War Medicine. 3: 484; 619. 1943.

somiasis was introduced into South America and Scott[3] reported in 1940 that *Schistosoma mansoni* was as prevalent and severe in some parts of Venezuela as any place in the world. In the males above 10 years the incidence of infection is about 90 per cent. This infection is readily acquired by contact with larvae in infected waters, and the Caucasian race is extremely susceptible. The British have a rule that if any soldier falls into fresh water in infected areas he must go immediately to a place where a disinfecting bath can be taken. The widespread migrations through areas where schistosomiasis is prevalent can result in some infections. Craig and Faust[4] speak of two cases after the last war of *S. haematobium* acquired in Australia from water contaminated by soldiers returned from the Middle East. Fortunately no outbreak occurred. Two cases of *S. haematobium* were also reported from Michigan by Blum and Lilga[5] in December, 1942, in boys who were originally infected in Africa. One of the cases was particularly interesting because the diagnosis was first made on a routine urine examination in April. But from the history it was certain that the boy became infected four months previously. In spite of the best quarantine precautions it is likely that this infection would have been missed when he entered the country in February. The schistosomiasis of the Old World has a close relationship with a schistosome dermatitis, commonly called "swimmer's itch," that occurs in the northern region of this country. The latter variety is relatively non-pathogenic for man, as it produces only a local reaction when the larvae penetrate and die in the skin. It is not known, however, that the more dreaded *S. mansoni*, *haematobium* or *japonicum* will not find snails here receptive to their development upon chance introduction. We know therefore that there are sufficient opportunities for our men to contract the various protozoan infections, especially malaria and dysentery which are already responsible for adding a large number of names to the casualty lists.

DISSEMINATION OF DISEASE BY MODERN TRANSPORTATION

Another epidemiological factor now confronting us for the first time is the speed of modern transportation, particularly by air. Not only is there an increase in speed but also in the amount of air traffic. It may be stated that this type of transportation has been going on for several years even through the heavily infected tropical belts. But there is a

[3] Scott, J. A. Schistosomiasis in irrigated mountain valleys of Venezuela. Am. Jour. Hyg., Sect. D, 31: 1. 1940.
[4] Craig, C. F., & Faust, E. C. Clinical parasitology, p. 371. 1940.
[5] Blum, B. B., & Lilga, H. V. Schistosomiasis infection: report of two cases found in northern Michigan. Jour. Am. Med. Assn. 121: 126. 1943.

difference now. Before the war, traffic was confined mostly to passengers of the higher income brackets who stopped at the better sanitated hotels in the larger cities and had a minimal contact with native population. Now most of the air travelers are individuals who have lived in close proximity with the infected natives, many for long periods. Actually air transport will be the means of returning large numbers of sick personnel, as it was announced in the press recently that air evacuation units are now in operation carrying 22 patients in each plane from Africa to the United States. Speed of travel is an important factor because it permits the entrance of infected individuals into new areas before the incubation period has elapsed. For example, in my own experience I recently arrived in the United States 34 hours after departing from Africa. It would have been possible to have contracted any of the infections previously cited shortly before leaving and yet have had several days of apparent good health in this country before diagnostic symptoms developed.

The task of excluding diseases and their vectors by the United States Public Health Service at regular ports of entry during normal times has been extremely difficult. Under disturbed conditions when large numbers of transoceanic planes are arriving, frequently unannounced and on irregular schedules at any one of scores of airports, the quarantine duties are multiplied enormously. With the exception of the route to the British Isles practically all of the foreign air traffic to the United States originates in the tropics—in the Southwest Pacific, South and Central America, the Caribbean area, Middle and North Africa, the Middle East, Russia, India and southern China. The airports in these places are usually located in the insanitated areas frequently hacked out of the jungle and surrounded by native villages whose occupants are being used as laborers. Every incoming flight presents a real hazard.

There is danger of transporting not only the diseases themselves but also their vectors. We have had a real lesson in what can happen from the accidental introduction of *Anopheles gambiae* into South America about 1930. This insect multiplied at a prodigious rate and the first six months of 1938 witnessed a malarial epidemic involving a minimum of 14,000 deaths and 100,000 cases. The eradication of this species of mosquito from Brazil by the joint efforts of the Brazilian government and the Rockefeller Foundation will go down in history as one of the outstanding achievements in preventive medicine.[6] The results of this campaign illustrate the necessity of applying energetic control methods. Neglect could have resulted in disaster for all of the tropical areas of the

[6] **Soper, F. L., & Wilson, D. B.** Species eradication. Jour. Nat. Malaria Soc. 1: 5. 1942.

Western Hemisphere. Incidentally, within the past year the same mosquito has been found on a few occasions in planes coming from Africa. Also one tsetse fly was found. Fortunately all insects were dead as the result of disinfestation of the planes with insecticides at their points of departure and arrival.

Insect vectors will be covered later in the symposium but their importance cannot be overemphasized. The volume of air transportation now being carried on is affording daily opportunities for bringing these vectors of disease into our country. Unless destroyed by spraying they will probably arrive in a healthy condition because of the brief time required for the journey. Finally, if they are able to establish themselves it is very probable they will multiply more rapidly than in their natural habitats. It is a fundamental law of nature that when any biological species is transported to a new favorable environment it multiplies very rapidly because it has left its natural enemies behind. Such examples are the Japanese beetle in the United States and *Anopheles gambiae* in Brazil.

A factor of major health importance in all wars is the development of epidemics among disturbed populations. Epidemics do occur in times of peace but they are usually confined to relatively limited areas because the topography or climate seems to check their spread. Last year I saw several thousand Poles coming out of Russia into Persia who had been uprooted from their homes and many of whom were ill. Many had virulent malaria acquired around the shores of the Caspian Sea. Several varieties of dysentery were evident in large numbers and cases of typhoid were extremely numerous. Typhus had just disappeared, as it was the onset of the hot season. In spite of aid they were receiving from the American Red Cross and the British they were living in a very makeshift fashion and they were contracting and spreading many diseases. There would be minimum danger except to themselves if they were to remain in this area, as there would be only an increase in the reservoir of the disease already present. The real danger will come as the result of their mass migration, as they soon will be dispersed into East Africa, India and other places. A migratory people can initiate epidemics in tropical areas even if the responsible disease is already prevalent. For example, it is a common occurrence for malaria to break out in troops when they are moved to other infected areas, even if they have previously acquired considerable tolerance to the disease. The explanation usually given is a lowered resistance due to fatigue, improper food, etc. A more likely reason is that they have come in contact with a strain of the disease with which they have had no previous experience. At the same

time they may carry with them a strain which will adversely affect the people who are permanent residents of the area. During the past war malaria returned home with the infected soldiers and secondary epidemics developed in several parts of England. At Emden, on the north coast of Germany, there were 6000 cases. The worst epidemic of all occurred in Russia where the starving civilians came down with malaria by the thousands. The Red Cross reported 3,000,000 cases in the Republic west of the Urals, in Georgia one-half of the population was affected, and in villages near Tiflis two-thirds of the population died of malaria. Malaria was present in all of these areas yet the probable introduction of new strains coupled with factors conducive to rapid transmission resulted in severe epidemics. The mass migration of returning soldiers can induce comparable consequences in this country.

SUMMARY

Summarizing the factors that are aiding the protozoan group of diseases and some of their near relatives in obtaining new footholds, we find the most important to be the massive concentration of susceptible troops in some of the foremost disease centers of the earth. It has already been found that these men are acquiring infections, particularly malaria and dysentery, in alarming proportions. Time alone will tell whether the other pathogenic agents can gain access to these fresh hosts and there is no reason to assume this will not happen as every opportunity is present. The rehabilitation of the sick will disseminate the acquired infections to every part of our country and it seems quite likely that there will be some favorable places that will serve as the foci for epidemics. We must be alert in recognizing the imported diseases in all of their stages so that they can be treated, isolated or otherwise controlled in order to minimize their danger to us. For the postwar period we will not only be concerned with the effects of tropical diseases on our own soil but will take a leading role in preventive medicine throughout the world. Training centers must be provided for increased teaching facilities and opportunities for fundamental research so that we can more ably fulfill the responsibilities that will come to us as the result of our greater contacts with these less familiar diseases.

Although in many respects the immediate prospects seem fraught with danger, I believe that the long-term picture is one of extreme opportunity of service for all mankind. Finally in this reorganized world that will surely follow the war, we cannot continue to follow the British, Dutch, Russians, Germans or any other nation in the study of tropical diseases but must assume a comparable role, if not a leading one.

DISCUSSION OF THE PAPER

Dr. W. H. Wright (*United States Public Health Service, Washington, D. C.*):

I am here in the difficult role of substituting for Dr. Thomas Parran, Surgeon General of the United States Public Health Service, who wishes me to express his regrets at his inability to be present at this Conference.

As the Federal agency most concerned with civilian health, the Public Health Service is not only interested in endemic disease but for 150 years has maintained a watchful eye lest the flotsam of some epidemic in some far-off corner of the globe be cast on our own shores. Now that we are engaged in a global war, the possibilities of the introduction and establishment of exotic diseases have been increased enormously. Consequently, the Public Health Service is vitally interested at this time in many diseases, protozoal and otherwise, which may become of domestic importance, or whose domestic distribution may be enhanced as a result of our military endeavors. Let us review briefly some of the present and postwar problems as they concern the public health significance of certain protozoal diseases.

Our malaria problem in World War I was confined to the southern camps and extra-cantonment areas. For the most part our troops were engaged in temperate or cold climates where malaria was not endemic. But now the sons of the veterans of the Marne and the Meuse, Archangel and Siberia, are in combat in some of the most highly malarious areas on the globe. Many of these men will return as carriers of the disease and many will go back to their homes in parts of the country which have long been free of the disease. A proper concentration of carriers in areas where there is a suitable concentration of vectors will lead to the establishment of new endemic centers of malaria. Furthermore, there is the possibility, if not the likelihood, of the introduction of new strains or new species in areas in which all the several species do not exist at present. We already have isolated examples of the potentialities of such occurrences even with the introduction of a limited number of carriers. For instance, Craig has cited the incident of the National Guard company from Connecticut, the members of which contracted *Plasmodium falciparum* infection in southern camps during the Spanish-American War and introduced this species in their home community on their return from the service, a community in which only *P. vivax* had previously been known. Matheson has cited the Aurora, Ohio, outbreak of 1934 to show the explosive effect of the introduction of a single case in a community, in which a suitable mosquito host was available. We may anticipate as a postwar development the probable occurrence of numerous instances of this sort.

On the other hand, our participation in the war will not be entirely on the debit side so far as malaria is concerned. Certain advantages will accrue to us. We must place on the credit side the gains we are making in the control of malaria in extra-cantonment areas in the South, our capacity to produce large quantities of synthetic antimalarial drugs, our research which should produce additional facts concerning the epidemiology of the disease, the ecology of vectors, and drug control, and last but not least the increased knowledge of malaria and other tropical diseases on the part of our physicians serving with the armed forces.

AMOEBIASIS

With amoebic dysentery also, we are faced with a situation somewhat different than we were in the first World War. Not so long ago, an eminent medical officer remarked to me that amoebic dysentery has never been a military problem except in the Philippines during the Spanish-American War and the Insurrection. His statement is entirely true but at the same time we must not forget that the Philippine campaign is the only one of consequence which we had fought in a tropical country. Now our troops are in combat in areas in which strains of *Endamoeba histolytica* are particularly virulent, not for the indigenes, but for susceptible individuals who have not developed resistance through previous and perhaps long-continued exposure.

The protection of troops against bacillary and amoebic dysentery is difficult under combat conditions. Superchlorination of water will destroy cysts of *E. histolytica* and portable filters such as used in advance zones will probably remove such cysts from water. However, even though the best sort of protection is provided, it

is not always possible under combat conditions to make use of available facilities. Consequently, the dysenteries must be reckoned with in any military campaign.

It is difficult to appraise the effect of the dispersal of numerous returning carriers of *E. histolytica* throughout our civilian population. The question was raised during the last war and Boeck and Stiles made a total of 13,043 examinations of 8029 individuals for intestinal parasites. These persons included overseas veterans, troops stationed in the United States, persons with no military service, and persons whose service connection was unknown. The incidence of *E. histolytica* in overseas soldiers was no higher than that encountered in the other groups. Conditions for our military operations are now entirely different and we may expect the return of a larger number of infected individuals at the end of this war. What effect these carriers will have on our civilian health is problematical but it is reasonable to assume that their dispersal may well lead to a higher morbidity rate from amoebiasis, and that perhaps new and more virulent strains may be introduced.

In the meantime, all our problems in amoebiasis are not postwar problems. Some have been with us for a long time. For instance, we are weak in our knowledge of the varying virulence of strains and of the factors influencing susceptibility to the disease; we need more potent and more specific antigens for serological and intradermal tests so that diagnosis can be simplified; we should have greater knowledge of the epidemiology of the disease and the relative importance of different modes of transmission; and our chemotherapeutic methods could stand a lot of improvement.

LEISHMANIASIS

Other than a few imported cases, leishmaniasis has not occurred in the United States. Even though the method of transmission of the visceral type of the disease has been definitely established by Swaminath, Shortt and Anderson, as occurring through the bite of *Phlebotomus*, we have difficulty in appraising postwar significance of the disease as a public health problem in the United States. We do know that protection against the vector is often impractical, if not impossible, and that in the present state of our knowledge only the most obvious cases are detected by routine diagnostic procedures. We may surmise therefore that individuals with the disease will return to the United States. I believe that only three species of *Phlebotomus* are known from the United States, two of them imperfectly. One species, *P. diabolicus*, occurring in Texas, is known to be a decided feeder on man. It is believed that studies should be conducted to determine the distribution and ecology of the indigenous species of *Phlebotomus* and that experiments are in order to determine whether *P. diabolicus* can serve as a vector of leishmaniasis.

TRYPANOSOMIASIS

The possibilities for the establishment of African sleeping sickness in this country are perhaps more remote than those in the case of some other exotic diseases. Of course, we do not have in this country species of *Glossina*, a fact which mitigates against the establishment of the disease in the continental United States. However, we do have other blood sucking flies including tabanids and *Stomoxys calcitrans*, the latter of which has been incriminated as one of the vectors of the disease. As African trypanosomiasis has exhibited no tendency to spread extensively in areas where *Glossina* does not abound, it would appear that the disease is unlikely to gain a foothold in areas where dependence on transmission is limited to other vectors.

Perhaps we should be more concerned over the possible establishment of Chagas' disease than of the African variety of trypanosomiasis. It is a well-known fact that naturally infected *Triatoma* have been found in several places in the South, Southwest, and California, and that reservoir hosts of *Trypanosoma cruzi* are present in this country. No human cases of the disease have apparently been found to-date. However, we are sending on different missions many individuals to endemic areas in Central and South America and furthermore we are importing Mexican labor to work in areas in which infected *Triatoma* have been found. We can only conjecture the effect of these various circumstances and hope that they will not be sufficiently fortuitous to bring about the introduction of human cases and the spread of the disease in this country.

One cannot leave this question without saying a brief word concerning the vigi-

lance necessary to prevent the introduction of disease vectors into this country. The reverberations which followed the introduction of *Anopheles gambiae* into Brazil are still sufficiently audible that no further example is needed to stress the importance of this sort of thing. The enormous expansion of air travel within the past year and the addition of transport routes to all corners of the globe have brought about a constant and increasing peril in this respect. The Public Health Service is alert to all of the potential possibilities in the situation and is exerting every effort to guard our shores against the introduction of disease-transmitting species. The adoption of aerosol spraying for airplanes has added to the efficiency of plane fumigation and we may expect the continued adoption of new plans and new methods as the need asserts itself.

Dr. Robert Matheson (*Cornell University, Ithaca, N. Y.*):

Some of my observations also illustrate the fact that malaria will remain in the blood stream for many years without necessarily causing symptoms. For instance, there is a record of a six-year-old child in Ithaca, N. Y., who developed a severe case of malaria (*Plasmodium falciparum*) within 48 hours of receiving a transfusion from her father, a Syrian, who had lived in the United States for 35 years without noticeable symptoms.

CHANGED VIEWPOINTS ON HELMINTHIC DISEASE:

WORLD WAR I VS. WORLD WAR II

By

Norman R. Stoll

Rockefeller Institute for Medical Research, Princeton, N. J.

It has seemed to me that of the several ways one might approach discussion of the helminthological problem in relation to American participation in the war, one way was to take stock of our general thinking in the field over the inter-world-war interval. This period has been a fruitful one. Its findings, added to and interpreted by earlier knowledge, obviously make the frame within which officers of the Medical Department, and particularly of the Sanitary Corps, of the Army of the United States, will take the actions designed to protect military personnel from helminthic adventures. It is hoped these selected viewpoints may light up angles which a more synoptic factual handling of the topic might not stimulate to the same degree. Whether or no, the pleasure I have in addressing you has had added to it the interest that has been afforded one returning for an hour to an examination of concepts in the field of the helminths of man.

If we were to take a global view of the helminthic problem—what, not too lucidly, could be called the globalminthic perspective—we might hazard the estimate that the world incidence was such as to represent the equivalent of every man, woman, and child, more than 2000 millions of them, being infected with parasitic worms. Half, say 1000 millions, may be presumed to harbor ascaris. Forty per cent, or some 800 millions, we will consider hookworm infected. The remaining 10 per cent, some 200 millions, would contain all the other dozens of species for which man is usually or occasionally host. This is not to say that every person is helminthically parasitized. Probably less than three-fourths are. It is of the nature of things helminthological that conditions permitting parasitization by one species may also permit additional species to be present. Parasitic organisms as large as worms are the shifting guests then of, say, a billion and a half people on the globe.[1] This is the hel-

[1] Let the point be emphasised that these are palpably estimates. No one seems to have undert a serious calculation of the helminthic incidence of the world population. More guardedly one say that between 500 and 1500 millions harbor ascaris, between 400 and 1200 millions bear hookw all the other species infecting between 100 and 300 millions; and that, if the minimal conjectur more nearly correct, due to multiple infections, only half the global population is parasitised by w

minthologists' backdrop of the war. One can go farther. Insofar as the professional helminthologist assumes that defaunation of the world's inhabitants approximates his responsibility in constructing an adequate civilization on the planet, it is a sobering challenge.

To be sure, these helminthic burdens are not distributed equally with respect to the land areas man inhabits—less evenly than is man himself. In the broad band of the tropics and subtropics two forces have helped to maintain a hyperendemic area. Not only is the parasite favored by environmental factors that tend to increase the opportunity to maintain the life cycle, and thus reach another host, but the human hosts themselves, live, let me say, in a closeness to nature, and a closeness to the threshold value of marked susceptibility, so that continuity of helminthic life cycles is distinctly favored. But the warm zones of the earth are not the exclusive homes of helminth-bearing human beings. The great diversities of their life cycles permit some species to maintain themselves in latitudes far from the equator, even in what we consider the cold northland. As this occurs in the in-between areas as well, then wherever American troops now are or may go it can be considered that helminths await them.

Let me illustrate. Several years ago in east central China we were conducting a survey in one of the missionary colleges. Its students were from homes on the average far above the low economic brackets, and only in minor degree were they the sons of farmers. In eliciting their interest in so strange a survey I predicted that every other one of them who sat in front of me harbored parasitic worms. If one as an individual did not, then the neighbors on each chair next to him did. It didn't turn out to be a bad guess, for 49 per cent were later proved positive. But that, you say, is China.

A few years later in the town of my present residence the occasion arose to examine a group of school children for helminths. Now you should know, if you are not already aware of it, that the line

"Fair Auburn, loveliest village of the plain"

could, out of all America, most properly be applied to Princeton, New Jersey. In so saying, I am thinking of many of the components of the good life that are supposed to represent what we are fighting for in this war. Yet we found the helminthic incidence in a suburban elementary school group, which was not regarded as a loaded sample, to be 23 per cent.

This is no reason for you to get nervous on your seats and wonder whether every fourth or eighth among you is suspect, although two pertinent incidents may be worth mentioning. A few years ago a dis-

tinguished non-helminthological professor of my acquaintance raised horrific reactions in his household when he displayed as a captive the adult *Ascaris lumbricoides* he discovered by chance he was failing longer to provide a happy home for; and a registered nurse in a Michigan city raised a howl over a similar event. These were trained observers. Even so they may not have been trained enough to know that the sleepless nights of their children were being incorrectly ascribed to many things psychological or microbic, except the nocturnal peregrinations of erstwhile lost *Enterobius vermicularis*, freshly come from the enteric regions.

Despite the greater concentration of our American population in the northeastern and north central states, I am inclined to believe that more than 30 millions of our own people bear helminths. When we take into account the striking incidence of *Trichinella* recent studies[2] have shown, and add to it an equally striking incidence of *Enterobius* recently improved methods[3] have permitted determining, both to be added to ascaris, hookworm, and tapeworms known to be in circulation, an over-all estimate of 25 per cent may well be conservative.

Now I realize that this is all elementary to an audience such as this; that what I have been saying—somewhat hazardously, I admit, as extrapolations always are—is merely to roughly quantify a picture with which you are already familiar. Why emphasize it? Partly to stress one of the changed viewpoints to which my title refers, that helminthic infections are not problems alone of distant tropical areas but of where we live, no matter where that may be. This, I think, is one of the contributions to our helminthological thinking that has been given a new emphasis in the quarter century roughly represented by the period between the two world wars.

As we face forward to the present war in contrast to those similarly facing forward to World War I, a number of other changed viewpoints based on new or fuller facts and techniques can be passed in review.

There is, for instance, a group of life histories, with the biological details of which we are now equipped, rather than with their generalities or suspicions. Earliest come, and perhaps of greatest consequence, are the *Schistosoma* life histories, which were so rapidly clarified[4] following confirmation of the Japanese work on their own blood fluke. After these came similar clarification for the Occident with respect to *Paragonimus*,[5]

[2] Hall, M. C. Pub. Health Rep., U. S. Pub. Health Serv. 53: 1472–1486. 1938.
[3] Cram, E. B., Jones, M. F., and Reardon, L. Rev. Med. Trop. y Parasitol. 7: 4–6. 1941. Hall, M. C. Livro Jubilar Prof. Travassos, pp. 195–211. 1938.
[4] Leiper, R. T. Trop. Dis. Bull. 20: 859–864. 1923. Faust, E. C., & Meleney, H. E. Am. Jour. Hyg. Monogr. Ser. No. 3: 339 pp. 1924.
[5] Nakagawa, K. Jour. Parasitol. 3: 39–43. 1916. Ando, R. Med. Woch., Tokyo. Nos. 2175, 4, 5. 1920. Ameel, D. J. Am. Jour. Hyg. 19: 279–317. 1934.

the lung fluke, *Clonorchis*,[6] the liver fluke, and *Metagonimus*,[7] *Fasciolopsis*,[8] and *Heterophyes*,[9] the intestinal flukes. Among the cestodes came the elucidation of the missing link in the *Diphyllobothrium latum*[10] story, the filling out of the picture with *Sparganum*,[11] and, recently, the filling in of the gap on the anoplocephaline, *Bertiella*[12]; and for the nematodes, *Onchocerca*.[13]

It was in this period, too, that we came to know[14] of an endemic *Onchocerca* area in Central America, where its ravages rival those in its African home. Demonstrated within our own north central states and southern Canada was endemic diphyllobothriasis[15]; a zone from which stretch long tentacles establishing endemic cases in distant cities like New York[16] and others with Jewish populations.

Perhaps one should also mention *Oxyuris incognita* and how its potentialities as a star were demoted to the stage of opéra bouffe by the explanation[17] of its occurrence. These life history and epidemiological facts each can render a military service now.

In another direction studies began on the correlation between chemical composition and anthelmintic efficiency.[18] These stimulated the exploration of halogenated hydrocarbons, and the consequent discovery by Hall and his co-workers of the anthelmintic efficacy of carbon tetrachloride[19] and tetrachlorethylene.[20] Later, this renewed interest also turned hexylresorcinol,[21] gentian violet,[22] and phenothiazine[23] to use against the intestinal worms of man. Of their present place we shall hear from Lt. Colonel Mackie.

In still another angle of the field there came to be more generally extended to helminth infections the practical utilization of serological and

---

[6] **Muto, M.** Verhandl. Japan. Path. Gesellsch. Tokyo 8: 151. 1918. **Faust, E. C., & Khaw, O. K.** Am. Jour. Hyg. Monogr. Ser. No. 8: 284 pp. 1927.

[7] **Yokogawa, S.** Centralbl. Bakt. I. Abt., Orig. 72: 158–170. 1913. **Muto, M.** Jour. Kyoto Med. Assn. 14: 15. 1917.

[8] **Nakagawa, K.** Jour. Parasitol. 8: 161–166. 1922. **Barlow, C. H.** Am. Jour. Hyg. Monogr. Ser. No. 4: 98 pp. 1925.

[9] **Cort, W. W., & Yokogawa, S.** Jour. Parasitol. 8: 66–69. 1921. **Faust, E. C., & Nishigori, M.** Jour. Parasitol. 13: 91–128. 1926. **Khalil, M.** Lancet 2: 537. 1933.

[10] **Janicki, C., & Rosen, F.** Bull. Soc. Neuchâtel. Sc. Nat. 42: 19–53. 1917.

[11] **Yoshida, S.** Jour. Parasitol. 3: 171–176. 1917. **Okumura, T.** Kitasato Arch. Exper. Med. 3: 190–197. 1919.

[12] **Stunkard, H. W.** Am. Jour. Trop. Med. 20: 305–333. 1940.

[13] **Blacklock, D. B.** Ann. Trop. Med. and Parasitol. 20: 1–48, 203–218. 1926.

[14] **Robles, R.** Bull. Soc. Path. Exot. 12: 442–460. 1919.

[15] **Vergeer, T.** Jour. Am. Med. Assn. 90: 673–678. 1928. **Magath, T. B.** Am. Jour. Trop. Med. 9: 17–48. 1929.

[16] **Waters, H. W., & O'Connor, F. W.** Jour. Am. Med. Assn. 99: 1941–1942. 1932.

[17] **Sandground, J.** Jour. Parasitol. 10: 92–94. 1923.

[18] **Caius, J., & Mhaskar, K. S.** Indian Jour. Med. Res. 7: 429–464. 1919; et seq.

[19] **Hall, M. C.** Jour. Agric. Res. 21: 157–175. 1921; Jour. Am. Med. Assn. 77: 1641–1643. 1921.

[20] **Hall, M. C., & Shillinger, J. E.** Am. Jour. Trop. Med. 5: 229–237. 1925.

[21] **Lamson, P. D., Ward, C. B., & Brown, H. W.** Proc. Soc. Exper. Biol. and Med. 27: 1017–1020. 1930.

[22] **Sieo, K. T.** Tr. Far Eastern Assn. Trop. Med. (7 Cong. Calcutta 1927) 3: 200–204. 1928. **De Langen, C. C.** Med. Dienst Volksgezondh.-Indië 17: 515–529. 1928. **Wright, W. H., Brady, F. J., & Bozicevich, J.** Proc. Helm. Soc. Washington 5: 5–7. 1938.

[23] **Harwood, P. D., Jerstad, A. C., & Swanson, L. E.** Jour. Parasitol. 24: Dec. Suppl.: 16. 1938. **Manson-Bahr, P.** Lancet 2: 808–809. 1940. **Bercovitz, Z., Page, R. C., & de Beer, E. J.** Jour. Am. Med. Assn. 122: 1006–1007. 1943.

skin tests. Notably have these been of interest and value in parenteral helminth infection, and the war may witness their wider use for assistance in diagnosis, especially for hydatid[24] and bilharziasis.[25]

In what is more distinctly its own field, namely the microscopic examination of feces for evidence of the presence of enteric infections, there has been very considerable development. Twenty-five years ago helminthologists and public health laboratory workers were still relying primarily on the technic of the fecal smear, extended to hookworm diagnosis by Grassi and the brothers Parona 40 years earlier,[26] although employed by Davaine[27] two decades before them. It's a rare worker who has not in recent years attached his name to some modified method of discovering what is helminthologically of interest[28] in human excreta. Almost inevitably these methods turned on rendering more readily visible to the microscopist more of the helminth eggs that might be present in an aliquot. It was obvious that that could not be the only item of interest about helminth eggs present in a fecal brew, and thus came procedures to determine readily how many were there. The medical officer in this war thus has a wider choice of technics and more knowledge of their individual usefulnesses and limitations than ever existed before.

These quests were stimulated primarily by the hookworm problem. On a world basis, hookworm is to be regarded as the most important helminthic pathogen, and one writer[29] on the iron-deficiency anemias believes that "in a world-wide census the commonest cause [of overt blood loss] would probably be . . . due to infestation by the hookworm." Perhaps in regard to hookworm one should go so far as to rate it as of more consequence than all other helminthic infections combined.

Something happened to hookworm this quarter century just past. Its attempted control in frontal, full-scale attacks throughout the world, on a pattern more inclusive than ever devised against any other helminth, has produced great gains. Of even greater significance, however, was what grew out of the disproportion that showed itself between the beautiful, theoretical simplicities of its problem, and the realistic complexities that arose when grappling with it under differing epidemiological conditions. Some of the factors involved deserve discussion because of their practical significance; others because they strike to

[24] Dennis, E. W. Jour. Parasitol. 23: 62–67. 1937.
[25] Fairley, N. H. Jour. Roy. Army Med. Corps. 32: 449–460. 1919. Culbertson, J. T., & Rose, H. M. Am. Jour. Hyg. 36: 311–315. 1942.
[26] Grassi, G. B., Parona, C., & Parona, E. Gazz. Med. Ital. Lomb. 38: 193–196. 1878.
[27] Davaine, C. J. Compt. rend. Soc. Biol. 4: 188–189. 1857.
[28] Lane, C. Trop. Dis. Bull. 30: 503–516. 1933. Stoll, N. R. *In* Damon: Food infections and food intoxications. Williams and Wilkins Co. Baltimore. Chapter 17: 253–290. 1928.
[29] Scott, R. B. Lancet 2: 549–552. 1932.

the heart of a bit of fundamental biology that still seems to me to be often neglected in thinking of helminthically parasitized hosts.

First as to the biological fundamental. As we consider disease caused by protozoa, bacteria, or viruses, we identify in our minds the diseased state with a large accumulation of the microbic agents involved. It does not matter for the present discussion whether the large number of microörganisms is there on account of their virulence, or enhanced virulence, or of a necessary degree of susceptibility of the host permitting their large accumulation. The height of disease is characterized either by excessive numbers of parasites, or some pathological condition induced by the recent presence of such numbers. Attention being largely centered on those which characteristically multiply within the host, a small inoculum, perhaps only a few or a few dozen or hundred bacteria, were all that was needed to assume that the host would become an *in vivo* culture of the organism.

This multiplication within the host is not the rule with helminths, however. Here accumulation of organisms is characteristically due to new arrivals from without. Failure to emphasize this fundamental distinction placed in the past too great an emphasis on the role of helminth infections of small size. Just as the helminthological taxonomist found in a beautiful macroscopic parasite something tangible upon which to erect a new species, so the physician, faced with evidence of helminths in his patient, straightway labelled the infection as a disease.

Just to illustrate. What happened in earlier hookworm control was that any person showing a hookworm egg was a subject for vermifuge. As a result we come upon records such as in Trinidad, for instance, where by 1921 it was reported[30] that of about 150,000 persons given anthelmintics, more than 1500 had been given 9 or more consecutive hookworm treatments in order to defaunate to the last ovipositing worm as gauged by the fecal examination methods used.

There was bound to be a reaction to this approach, for besides imposing practical consequences such as to render hookworm control impossible on the deservedly large scale the need implied, it imposed a chemotherapeutic burden on the individual that tended to rival the damage of the helminthic burden. Both of these ideas have relevance for the problem the medical officer faces in war time.

A reaction came. One sign of it was an article[31] in 1916 by a public health worker in the Far East, entitled, "Are there harmful and harmless

[30] Payne, G. C. Report on work for the relief and control of hookworm disease in Trinidad, from August 11, 1914, to December 31, 1920. International Health Board, Rockefeller Foundation, New York. 31 pp. 1921.

[31] Heiser, V. G. , Jour. Sociolog. Med. 17: 37–46. 1916.

hookworm infections?", with observations tending to support the obvious answer to a rhetorical question. The Uncinariasis Commission, which worked in the Orient from 1915 to 1917, established the distinction with adequate evidence,[32] using the method of vermicidal treatment and recovery of worms expelled. The point of view was well summarized by Darling[33] later when he divided "cases infected with hookworms into three groups, based on the degree to which compensation for blood losses occurs: Group A. Blood loss compensated. Group B. Compensation disturbed and breaking. Group C. Compensation broken." Of his Group A he observed: "Most of these people carry less than one hundred hookworms, though a few may harbor as many as 200 or 300."

Later came the accession of a simplified indirect method of determining relative worm burdens by means of dilution egg counting.[34] Once the possibility was open of testing the general idea it was confirmed by various workers, as for instance in China,[35] in southern Alabama,[36] in Mexico,[37] and in the Argentine.[38] Some of these studies added an incidental observation largely lost in the general problem, that gives added point to Darling's Group A. This evidence[36] was that persons lightly infected with hookworms had slightly higher average hemoglobins than the negatives—direct support of Darling's idea of compensatory effect.

In these days when blood donors contribute a pint periodically for war purposes, some of them as frequently as once in two months, there is perhaps less need of becoming disturbed over the exaggerated blood loss in a light hookworm infection. Its stimulation of the hematopoietic mechanism may more soberly be balanced against the known treatment risk with even the best of our anthelmintic chemicals. And the best of these, be it remembered, should dislodge 3 to 9 times as many parasites in the first treatment, as will all subsequent treatments combined.

Evaluation of this state of affairs in connection with hookworm has been aided by the fact that this bloodletting parasite has permitted an objective registration of its damage in lowered hemoglobin readings. Most of the helminths of man give rise to no such simple quantitative

[32] **Darling, S. T., Barber, M. A., & Hacker, H. P.** Hookworm and malaria research in Malaya, Java and the Fiji Islands. Report of Uncinariasis Commission to the Orient, 1915–1917. International Health Board, Rockefeller Foundation, New York. 191 pp. 1920.

[33] **Darling, S. T.** Nelson loose-leaf medicine. Chapter VI, pp. 477–489. 1922.

[34] **Stoll, Norman R.** Am. Jour. Hyg. Monogr. Ser. No. 9: 30–58. 1929. **Stoll, N. R., Cort, W. W., & Kwei, W. S.** Jour. Parasitol. 13: 166–172. 1927. **Sweet, W. C.** Jour. Parasitol. 12: 39–42. 1925. **Brown, H. W., & Cort, W. W.** Jour. Parasitol. 14: 88–90. 1927. **Soper, F. L.** Am. Jour. Hyg. 7: 542–560. 1927. **Augustine, D. L., Helmy, M., Nazmi, M., & McGavran, G.** Jour. Parasitol. 15: 45–51. 1928.

[35] **Stoll, N. R., & Tseng, H. W.** Am. Jour. Hyg. 5: 536–552. 1925.

[36] **Smillie, W. G., & Augustine, D. L.** Am. Jour. Dis. Children 31: 151–168. 1926.

[37] **Carr, H. P.** Am. Jour. Hyg. 6: July Suppl.: 42–61. 1926.

[38] **Fülleborn, F., Dios, R. L., & Zuccarini, J. A.** Arch. Schiffs- u. Trop. Hyg. 32: 441–461. 1928. **Fülleborn, F.** Brit. Med. Jour., pp. 755–759. 1929.

estimate of the damage they do. It is worth bearing in mind, however, that the basic assumptions should hold good.

This is not to say that a single parasite of a given species may be unable to produce some degree of host reaction, a degree varying with the particular worm species and the stage of susceptibility of the host. It is to say, however, that the pathogenic hazard occasioned by a few helminths may be of the order of pinpricks or scratches with which the host is biologically accustomed to deal without jeopardy.[39] We might, as a working basis, define helminthic disease as being the level of host-parasite relation at which the helminthic accumulation places the host in jeopardy of its well-being or its life. Or one might say that helminthic disease is at least a shade more than is implied by the Christian Science spelling of the word in its hyphenated form, as dis-ease. Either criterion gives the physician the opportunity of proceeding with the case before him in terms clinically realistic and helminthologically sound. If his evaluation of a sympton complex makes routine use of such assistance as, for instance, a hemoglobin determination, may he not be under an equal responsibility to get some objective measure where possible of the size of the helminthic infection before him? For when there is no multiplication of the parasite within the host, microscopical demonstration of the presence of an intestinal helminth is not equivalent to demonstration of a diseased state. Undue emphasis on the microscopic positive may indeed lure attention from other causes of greater etiological significance. This has occurred with the lay physician; it could also occur with the army or navy medical officer.

If this changed viewpoint, which may be called the recognition of the third dimension in helminthic infection, has clarified our thinking in certain directions, it is appropriate to state that it has led us into new points of view in others.

You may, for instance, recall the report[40] a few years ago of the fate of a group of helminthic infections studied over a period of 15 months, in the absence of reinfection and also in the absence of treatment. In these children in a Tennessee institution all of 22 ascaris cases became egg-count negative, as did 11 of 14 whipworm, 10 of 12 *Hymenolepis nana*, and 4 of 7 hookworm infections. No explanation of occasionally missing a diagnostic egg could obscure that something of interest was happening here. And in the last war, one author[41] referring to troops in the Middle East wrote: ". . . the longer their service in Mesopotamia, the greater the tendency for a natural cure" of helminthic infections. In

[39] Le Bas, G. Z. L. Jour. Helm. 2: 151–166. 1924.
[40] Keller, A. E. Jour. Am. Med. Assn. 97: 1299–1300. 1931.
[41] Acton, H. W. Indian Jour. Med. Res. 6: 601–613. 1919.

both instances the writers emphasized the good sanitary conditions preventing reinfection. They might equally have observed that perhaps nutritional conditions were involved.

May I mention two further revealing sets of facts? To one distinguished American parasitologist, a microscopic positive represented hookworm disease. Any quantitative method was a Euclidean red herring. Hookworm disease presented so striking a picture that one could go into a schoolroom and by "streetcar diagnosis"[42] say there, and there, and there, was hookworm, and get it verified by fecal examination. It is interesting that Stiles, when crowded by the argument in oral discussion, was willing to assert that if the symptom complex was not verified by the microscope it was hookworm disease just the same, eggs or no eggs. In his enthusiasm against what he considered an oversimplified arithmetic point of view, Stiles and others have overdrawn, I believe, the position they opposed. Certainly not in my contacts with anti-hookworm workers has the idea been current that if a given patient with low egg count presented a clinical picture of hookworm disease he should not be appropriately treated as an individual case.

Before drawing a further conclusion on Stiles' point of view, let us consider briefly some findings from up-country Panama. We had become familiar, as noted a few paragraphs earlier, with the idea that hemoglobins and worm burdens as represented by gg counts were in high inverse correlation with each other. That is, the egg count and degree of anemia tended to be in parallel. In some of the Panama groups this was not true. In the words of the report[43] concerning one district: "There is . . . no negative correlation between the individual hemoglobin percentages and the size of the egg counts. . . . This is very astonishing when the . . . intensity is considered." Here, in host groups widely separated geographically and measured by very different means, is evidence of a state of affairs that has come to play an increasing role in the determination of the host-parasite relation. In the nature of the case it is a situation that has been explored primarily in experimental animals. I refer to the general conception that the condition of the host is a significant factor in determining what the

[42] **Stiles, C. W.** Jour. Am Med. Assn **98**: 2189–2190. 1932
As Dr. V. G. Heiser emphasized, discussion of one of Dr. Stiles' points of view in no wise compromises the extraordinarily great contribution which Dr Stiles made to an adequate comprehension of the hookworm problem, and the practical development of measures to control it. He was personally the greatest single American force in promoting these factors, which found expression in the establishment in 1909 of the Rockefeller Sanitary Commission for the Eradication of Hookworm Disease. This Commission became, in a sense, the progenitor of the International Health Division of the Rockefeller Foundation, which itself, in the period 1913–1937, contributed $3,846,324.88 to hookworm control (and investigation) throughout the world. One of Dr. Stiles' last articles, entitled, "Early history, in part esoteric, of the hookworm (Uncinariasis) campaign in our southern United States" (Jour. Parasitol. **25**: 283 306. 1939) will be found of special interest, even to the lay reader.
[43] **Cort, W. W., Schapiro, L., Sweet, W. C., Stoll, N. R., & Riley, W. A.** Am. Jour. Hyg. Monogr. Ser. No. 9: 139–160. 1929.

result of infection will be. Hosts are not of uniform susceptibility to hookworm. How trite that sounds once you state it, and how long we were in taking it into account! As illustrations with bloodletting helminths, in both sheep with *Haemonchus contortus* (a blood sucker of hookworm caliber) and dogs with *Ancylostoma caninum*, it has been possible to demonstrate host effects varying from almost complete insusceptibility induced through previous infection—immunity, if you will—to increased susceptibility due to decreased stamina of the host produced, for instance, through malnutrition. Especially in the present connection are some of the Baltimore experiments[44] of interest, wherein the state of immunity to dogs could be rendered labile through manipulation of the diet. However, the experiments proceeded only so far as to demonstrate a marked correlation between poor nutrition and increased susceptibility, as compared to animals on the routine laboratory diet, which by contrast was considered "adequate." We need, further, the positive data, of what specific dietary constituents can assist a host in developing immunity or maintaining resistance. Some results may be furnishing clues to these positive factors. Especially of interest are studies[45] on the relation of vitamin A to infections developing in chickens with *Ascaridia* and with *Heterakis*, and in rats with *Trichinella* and with *Strongyloides*. Also studies[45] on vitamin B in relation to infections developing in chickens with *Ascaridia*, and more recently, in a cognate field, experiments[46] indicating increased resistance to malaria in fowl given adequate amounts of biotin. A high protein diet has been shown[47] to counteract the ill effects of tapeworm infection in chickens.

Nevertheless, one may go back now to the streetcar diagnosis of Stiles and well believe that the affirmation of its being "hookworm disease" without demonstration of the organism was correct, for it was something more than simon-pure hookworm to begin with.[48] (As some of the patients diagnosed for Noguchi to study in South America were something more than simon-pure yellow fever.) And whatever tentative explanation was put forward, using race as a crutch, for the apparent

[44] Foster, A. O., & Cort, W. W. Am. Jour. Hyg. 16: 241–265. 1932; 21: 302–318. 1935.
[45] Ackert, J. E. Jour. Parasitol. 28: 1–24. 1942.
[46] Trager, W. Science 97: 206–207. 1943.
[47] Luttermoser, G. W., & Allen, R. W. Poultry Sc. 21: 111–115. 1942.
[48] Pertinent here are Bailey K. Ashford's remarks in his autobiography ("A soldier in science." 1934. Wm. Morrow & Co. New York): "For the benefit of intelligent readers, both lay and professional, both here and abroad, we amply acknowledge the very important place taken by nutritional unbalance in producing the picture recognised all over Puerto Rico as *la anemia*, or, in other words, hookworm disease. But the undeniable fact remains that this was not a nutritional, but a parasitic anemia. That it was capable of cure and of prevention by specific drugs and the use of latrines, respectively, is seen from the history of the sensational transformation of a helpless anemic at death's door to a ruddy vigorous laborer, simply on the expulsion of these tiny worms, and without any alteration in his accustomed food. How, then, does the influence of poor diet come in? Thus: parasites plus poorly balanced food bring fatalities and serious grades of anemia, which would not occur from parasites alone. . . . Only, however, when the fundamental thing is done will the [hookworm] disease, as a disease, disappear from Puerto Rico. And that fundamental thing is the provision of better food for Puerto Ricans, and much more meat. It is, therefore, no longer a medical problem, but a sociologic one of the very first water."

discrepancies of the Panama hookworm findings, the fact that low hemoglobins were found with low egg counts may have been the same phenomenon Stiles encountered; while high hemoglobins with high egg counts may equally have been a manifestation of an immune state, showing, as it not infrequently does in experimental animals, a raising of hemoglobin *before* the casting out of a resident infection.

With the new and deservedly great emphasis on nutrition, not in terms of calories alone but also in terms of adequacy in accessory vitamin and inorganic components such as iron, we may find ourselves some day just turning the hookworm picture around. We may then come to conceive of hookworms not as gross pathogens able at will to "muscle" their way around in the human host, for which the medical consultant can do nothing except clear out the worms, and prevent exposure to new infection. We may instead look upon those worms more as scavengers, demonstrating their ability to get the upper hand of hosts already prepared for their reception by a depleted state nutritionally or otherwise. For another side of the picture is as striking as classical hookworm, namely the degree of protection that well-fed hosts are able to develop through an immunity to hookworms and hookworm-like parasites.[49]

You properly ask, if an immunity is part of the functional phase of the host–parasite relation between man and his hookworms, why do we not know more about it, and how can we observe it at work in the world? One answer is that we do not yet have a serviceable approach, serological or otherwise, except the result of the reinfection itself, by which we can diagnose a state of immunity against hookworms, or probably of any helminth, although signs are increasing that we may get one. Another answer is that the terms in which we do gather our information on hookworm infection have not permitted the analysis of our data to reveal it. The latter point I wish to illustrate by comparing the average egg counts on a group of infected animals under conditions of continual reinfection. Although half the animals were immune and showed low egg counts throughout the period of observation, and the other half were demonstrating stages of susceptibility followed by immunity acquired through exposure to the infection, a single average eggs-per-gram count for the group at any time would not have revealed this Periodical examinations, properly analyzed, might accomplish this.

Without attempting too much in the way of prediction, I should at least like to give my opinion that the lecturer who surveys the changed viewpoints on helminthic diseases, World War II vs. World War III, will give a prominent place not only to the more specific role of nutri-

[49] Stoll, Norman R. Am. Jour. Hyg. 10: 384–418. 1929; 16: 783–797. 1932.

tional components in resisting hookworm infection but to technical procedures that will diagnose the presence of an anti-hookworm immunity in a given host, and the prescription for specifically inducing it if absent. Nor will hookworm be the only helminth concerning which he will be able to discuss such results. There are Russian reports,[50] as you know, which raise the presumption in favor of an immunity developing against *Diphyllobothrium latum*. This prompts the natural query as to whether the still obscure puzzle[39] of broad tapeworm anemia may not be the result of the combination of *D. latum* and an improperly nourished host.

On this general problem the war may well reveal new information of value. With the care given to provide the American soldier with an adequate diet, helminths he encounters are going to meet hosts better nourished and thus biologically better equipped to deal with the invader than is usually the case. But this will not always be true. For the rigors of military necessity will provide many chances for the opposite state of affairs. Medical officers may well have opportunities during the present epic, in which the nutritional cue will be of significant value during the emergency, and provide new clues for postwar experiment and rationalization.

There is another changed viewpoint which deserves mention now because some accident may permit observations of interest upon it during the extraordinary laboratory study that military personnel under the stress of war presents. Without too much flippancy I wish to refer to it as the problem of helminthic quislings—ordinary citizens who represent the enemy. In this case the citizens are presumably familiar helminths who may be harboring and transmitting virus diseases.

Those of you who have been following the virus field know of the now recognized etiological relationship of a virus and influenza. And all of you know that influenza under epidemic conditions[51] has its spread readily explained by droplet or contact infection. What is less well recognized is the peculiar problem of inter-epidemic carryover not only of influenza but of some other virus diseases. Recent work[52] has shown that in the case of swine influenza, the reservoir and vector, if you will, of "hog flu" from year to year is the larval stage of the swine lung worm, *Metastrongylus*. The nematodes, living as adults in the lungs of swine, produce eggs, which characteristically reach the outer world in the fecal droppings, where they are ingested by earthworms in which the infective larvae develop. Earthworms containing these intermediate stages, upon ingestion by hogs, are thereby enabled to convey the infective larvae to

[50] Tarassov, V. Ann. Parasitol. Hum. et Comp. 15: 524–528. 1937.
[51] Francis, T. Science 97: 229–235. 1943.
[52] Shope, R. E. Jour. Exper. Med. 74: 49–68. 1941; 77: 111–126; 127–138. 1943.

new hosts, in which the parasitic larvae wander from the digestive tract to the lungs and establish themselves. It is now known that *Metastrongylus* larvae developing from eggs shed by worms resident in a hog suffering from influenza are able to convey the virus, retain it for months (including overwinter), through the earthworm stage and, as infective larvae, transmit the virus in turn to susceptible hogs which they later come to infect. There is an oddity or two in the story which should not deter its acceptance. It has not been possible to free the virus directly from the lungworm larvae by any methods yet tried, and it may therefore be in a masked form. Nor do such larvae free the virus within the swine except under conditions when the swine host is in some state of physiological shock. Both experimental and field observations are in harmony, however, that this is the efficient method of transmission of "hog flu" from one season to another. These nematodes thus function as intermediate hosts of swine influenza.

Consider well, then, that part of the puzzle of epidemiology, perhaps especially of virus diseases, in which disease agents hide during interepidemic periods and of how new epidemics get started. May the state of affairs illustrated by swine influenza and *Metastrongylus* be duplicated in essential aspects by such virus diseases as poliomyelitis and the biological cooperation of some helminth of man which serves it? Does this cause one to think of possibilities in the wandering of *Ascaris*, of *Trichinella*, of *Strongyloides*, or even of hookworm? It can be left to the future to confirm or deny whether such an incredulous state of affairs obtains. In any event here is another hint of how closely helminths are tied in to some of the central problems of pathology.

Aside from these more general problems, what, you ask, is to be said more specifically regarding helminths in relation to the present dispersal of American troops?

In the British Isles and northward one would say the situation is more or less the equivalent to that here at home.

In northern Africa and the Near East the chief additional complicating factor is *Schistosoma*. It is rather interesting that a hasty survey of the entozoon records of troops in the Near East during the last war shows[53] principally hookworm, ascaris, whipworm, and schistosome infection We know that Algeria and Tunisia[54] have endemic *S. haematobium*, that it is not absent from Tripolitania and Cyrenaica, and that besides its great concentration in Egypt, it reaches on into Asia Minor, although it

[53] **Bardachzi, F., & Barabas, Z.** Münch Med Woch **64**: 570–572. 1917 **Fricke, W.** Deut. Med. Woch **43**: 845–847. 1917. **O'Connor, F. W.** Jour. Trop. Med & Hyg. **22**: 166–167. 1919. **Stewart, F. H.** Brit. Med. Jour, p 592. 1920

[54] **Baugé, R.** Arch. Inst. Pasteur Tunis **30**: 291–301. 1941. **Alcay L., Morill, F. G., & Musso, J. C.** Arch. Inst. Pasteur d'Alger. **20**: 89–99. 1942.

does not go as far as India. Of course beyond that we enter the *S. japonicum* area. One report[55] concerning bilharziasis during the 1916 Palestine campaign merits special mention: "When the Northamptonshire battalion had arrived at Kubri, men were warned, both on parade and in orders, of the danger of bathing in the Sweet-water, and this order was republished monthly in orders. Nevertheless nearly eighty men in the battalion bathed on at least one occasion in the Sweet-water Canal, and nineteen of these men subsequently developed vesical bilharziosis." In addition to bathing and swimming, soldiers may obviously expose themselves by wading, fishing, and washing clothes in cercarial-infested waters. Some of you recall the cases of British and American naval officers and seamen on the Yangtze patrol who became infected with *S. japonicum* after hunting trips in the neighborhood of that river.[56]

We can anticipate that there will be plenty of work for the parasitologists of the Sanitary Corps in areas of known or suspected schistosomiasis, especially where troops are to be garrisoned. The risk rate of untreated waters will need to be evaluated in terms of the amount of schistosomiasis in the inhabitants of the occupied area, and the local abundance of the appropriate snail intermediate hosts. Some chemical and subsidiary measures, such as the use[57] of a steam jet in weedy patches, may assist in reducing snail populations, and free chlorine (introduced, for instance, as chloramine[58]) may be used for cercaricidal purposes in containers or small bodies of water. In no tasks assigned to them is it likely that the parasitologically trained personnel will have heavier responsibilities than in determining with promptness and all necessary efficiency the location of snail foci to be avoided, and especially the securing of the necessary cooperation of troops in preventing exposure to infection, particularly when off duty. It is imperative to apply the admonition: "With the information at the disposal of troops, bilharziosis should now be treated as one of those diseases for which the individual is mainly, if not entirely, personally responsible."[59] And evidently in the last war orders were not sufficient.

In the Near East and India the problem of dracontiasis will obtrude itself. While its control is simple, by merely not drinking *Cyclops*-contaminated water, the long delay in the appearance beneath the skin

[55] Searle, C. Jour. Roy. Army Med. Corps 34: 15–34. 1920.
[56] Lambert, A. C. Tr. Soc. Trop. Med. and Hyg. 3: 278–302. 1910; 5: 38–45. 1911. Laning, R. H. U. S. Naval Med. Bull. 8: 16–36. 1914.
[57] Miyajima, M. Jour. Pub. Health Assn. Japan 14: 1–6. 1938.
[58] Blackmore, M. S. Jour. Roy. Army Med. Corps 51: 262–264. 1928. Witenberg, G., & Yofe, J. Tr. Roy. Soc. Trop. Med. and Hyg. 31: 549–570. 1938. Sproule, J. C. Jour. Roy. Army Med. Corps 73: 384–388. 1939. Magath, T. B. U. S. Naval Med. Bull. 40: 287–288. 1942. Braune, J. F. Deut. Trop. Z. 46: 409–426. 1942.
[59] Leiper, R. T. Jour. Roy. Army Med. Corps 30: 235–260. 1918.

of adult *Dracunculus* may render somewhat ineffective as a deterrent the value of bad examples among careless personnel. One thinks again especially of the factor of men heedless when off duty.

Among troops in the Orient and the islands one sees the complication of filariasis, as Prof. Matheson will emphasize. Besides the skin-penetrating *Schistosoma* in this region too, there is a group of helminths the protection against which is care in avoiding insufficiently cooked food. Kitchen practice in the American army mess is probably adequate to protect against the helminths passively transferred on locally secured vegetables, in meat and in fish. Several years ago the British army, following the appearance of what seemed an unusual number of cases of cysticercosis, conducted tests to determine whether the cooking, frying and roasting temperatures of pork products in army kitchens were sufficient to kill any cysticerci present. Their conclusions[60] were affirmative, and would equally apply to *Trichinella*, as well as to metacercariae in fish. One is inclined to believe that the presence of such infections may be principally ascribed to exposure off post, in the sampling of native dishes, etc. Members of the armed forces in foreign lands may be visualized as ever-curious travelers seeking to sample the exotic, when out of the line of duty. If so, the precaution appropriate to the lay traveler is in order—not to chew but to eschew raw native vegetables, and raw, pickled and smoked native meat and fish products, and relishes. That such precautions may be highly protective can be adduced from the experience of Occidentals living in the heart of the Orient, months on end, and in close contact with the inhabitants, without contracting more helminths than possibly a few hookworms through a cracked shoe. There is evidence that the Medical Department has points of this kind in mind. In the recently described fishing tackle kits for use in emergencies which are supplied to life boats, life rafts and floats, the waterproof instructions accompanying them include the statements[61]: "The flesh of fish caught in the open sea is good to eat, cooked or raw . . . Freshwater fish of any kind . . . are all unsafe to eat unless thoroughly cooked."

Besides helminthic infections, the presence of which in native populations represents a threat to outpost troops, there is one animal reservoir for which the helminthologist needs to express special concern. I refer to the dog bearing *Echinococcus granulosus*. From New Zealand to Australia to India to parts of the Mediterranean littoral and beyond, hydatid is not to be taken casually. Reports from autopsy surveys in

[60] Jour. Roy. Army Med. Corps 64: 92–100. 1935.
[61] Stillman, D. N. Y. Herald-Tribune 3: 3. March 7, 1943.

India show 28 per cent of pariah dogs examined at Rajanpur in the Punjab and 10 per cent of city dogs in Calcutta harbor *Echinococcus*, while nearly one-third of the country dogs in New Zealand are estimated to be similarly infected.[62] An indirect measure of the potentialities of the reservoir is the 1935 New Zealand estimate[63] "that at the present moment at least ten millions of our sheep have hydatid cysts in their livers." The peculiarities of this infection, in which man, like the sheep, acts as the intermediate host, need no emphasis to this audience, but one believes could well become the subject of a special pamphlet for every soldier on foreign service.

In regard to army sanitary standards in routine fecal disposal, the helminthologist need not express marked apprehension. Concentrations of feces, with or without urine, we now know[64] tend to be self-sterilizing for helminths, although fly access to recent accretions would be considered a constant threat to the dissemination of enteric infections with pathogenic bacteria and protozoa. The helminthic dangers from fecal deposits arise not from the use of latrines, but from the soiling of the latrine area where worm stages may secure sufficient oxygen and moisture to proceed with their development. Shallow trenching of infested excreta becomes a similar threat, and one which is increased by the secondary possibilities of dispersal that arise from deluges of rain in warm climates.

Similarly the more recent emphasis on the possibilities of air-borne[65] infection of infective helminthic eggs can as yet scarcely be assumed to have more than academic interest in the grand strategy of helminths vs. man.

How finally should one evaluate the possibilities of new postwar helminthic invasions to the United States? My inclination is to regard them as not representing a major public health threat. Perhaps most to be feared would be the establishment in the South of snail hosts which *Schistosoma* could utilize. The malacologists, I believe, have not yet encountered the preferred species there. Nor has *S. mansoni*, present in Puerto Rico and other outer parts of the Caribbean littoral, become established in continental U. S. A. In addition to that fact is the great experiment conducted by the importation in other centuries of African slaves. These must have brought both *Schistosoma haematobium* and *mansoni* as well as other helminths unaccustomed to the American

[62] **Sami, M. A.** Indian Med. Gaz. 73: 90–94. 1938. **Maplestone, P. A., & Bhaduri, N. V.** Indian Jour. Med. Res. 28: 595–604. 1940. **Barnett, L.** New Zealand Med. Jour. 40: 273–278 1941.

[63] **Barnett, L.** New Zealand Med. Jour. 34: 258–260. 1935.

[64] **Oldt, F.** Am. Jour. Hyg. Monogr. Ser. No. 7: 265–291. 1926. **Stoll, Norman R.** Am. Jour. Hyg. Monogr. Ser. No. 7: 292–379 1926. **Stiles, C. W.** Jour. Parasitol. 13: 47–55. 1926.

[65] **Nolan, M. O., & Reardon, L.** Jour. Parasitol. 25: 173–177. 1939.

scene, on a scale and under conditions vastly more foreboding than will returning soldiers. It is true we thus gained *Necator*, but seem not surely to have acquired other helminths of man thereby. A natural experiment of similar import has been commented on by a French worker[66] concerned with the possibility of schistosomiasis becoming established in France as the result of introduction by infected African soldiers. No autochthonous case is known, although infected persons have entered France for more than 100 years, many of Napoleon's soldiers having returned with haematuria.

The worms which especially typify the Orient are flukes, whose life histories require peculiarities of food habits, especially in regard to fish. It seems unlikely that a more threatening case can be made out for these in the postwar period than was made out for the earlier denied danger of *Clonorchis* becoming established by infected immigrants.[67] Nor do the possibilities of *Paragonimus* seem to be increased, even remembering our reservoir of appropriate intermediate hosts in the Central West.[68] Unless the return home of our military men should result in widespread changes in food[69] and sanitary habits, which seems less than likely, even the presence of proper intermediate hosts is a matter for public health awareness and caution rather than of fear.

Several years ago a prominent English helminthologist[70] reviewed "present day teachings on helminthology in relation to public health." He mentioned that "the intense activity with which research has been pursued during recent years has resulted in knowledge, which, if properly applied, ought to ensure the control and eventual eradication of the infestations of man with parasitic worms," but, he added, "on these questions the medical officer of health is as a rule lamentably ignorant."

As the helminthic aspect of American participation in the war and postwar can, in synopsis, be considered a specialized problem in public health, I wish to quote further from his "general summary of ascertained facts": "From the public health standpoint the following are the significant facts concerning helminth transmission in general

"1. The parasitic worms do not multiply within the human body. . . .

"2. The eggs or embryos must first leave the human body which harbours the parents. . . .

"3. The eggs or embryos are not infective to man at the moment of leaving the human body. . . .

---

[66] **Peltier, M.** Rev. Prat. Malad Pays Chauds. 9 253 260, 263–265 1929

[67] **Faust, E. C.** Nat. Med. Jour. China 9: 342–345. 1923. **Wayson, N. E.** Pub. Health Rep., U. S. Pub. Health Serv. 42: 3129–3135 1928

[68] **LaRue, G. R., & Ameel, D. J.** Jour Parasitol 23: 382–388 1937.

[69] **Stoll, Norman R.** *In* Damon. Food infections and food intoxications. Williams and Wilkins Co. Baltimore. Chapter 16: 219–252. 1928.

[70] **Leiper, R. T.** Brit. Med. Jour., pp. 110–113. July 19, 1924.

"4. Developmental changes occur during the period of delay outside the body and are essential to the formation of the 'infective' stage. . . .

"5. The environmental conditions requisite for the developmental changes outside the body vary with different species. . . .

"6. The parasite in its infective stage enters the body in most cases by the mouth in food or as contaminations of food, but in certain instances actively pierces the skin. . . .

"7. After entry in the infective stage the parasite has, in many species, to undertake an extensive pilgrimage within the body. . . .

"8. Few of the parasitic worms are specific to man. . . .

"9. The spread of helminth infection can be controlled by breaking the life-cycle at certain vulnerable points."

There is no need to impute any lack of awareness of such points to the Medical Department of the Army of the United States, as it deals with its helminthological front. The activity of its parasitology training centers testifies rather to foresight and preparedness. And one may hope, quite without presumption, that this occasion, fostering discussion of changed as well as unchanged viewpoints, will stand in a contributory relation to the solution of the specialized problems the field reveals.

DISCUSSION OF THE PAPER

Dr. D. L. Augustine (*Harvard Medical School, Cambridge, Mass.*):

Postwar helminth problems need not be greatly feared, but precautionary measures should be taken against the entry of "Old World" hookworm which is more dangerous in every way than our own hookworm. Furthermore, the development of elephantiasis in 20 weeks in some of our troops indicates that our men lack an immunity possessed by the natives, in whom the disease does not develop until the 50–60 year age group.

Dr. G. L. Graham (*Rockefeller Institute for Medical Research, Princeton, N. J.*):

The relation between diet and immunity has been clearly demonstrated; immunity may be lost when the diet is deficient. Since some dietary deficiencies do not break the immunity, it is possible it may be sustained by some other factor.

Dr. W. H. Wright (*United States Public Health Service, Washington, D. C.*):

Filariasis may not be a major public health problem, but it certainly must not be dismissed as a minor one, considering the ease with which it can be brought into the country.

Dr. V. G. Heiser (*New York, N. Y.*):

Dr. Stoll's egg-count method has been of great value in the efforts of the Rockefeller Foundation to stamp out helminthic disease in the Orient, especially in the economy of limiting treatment to the heavily infested areas.

ARTHROPODS AS VECTORS OF HUMAN DISEASES WITH SPECIAL REFERENCE TO THE PRESENT WAR

By Robert Matheson
Cornell University, Ithaca, N. Y.

INTRODUCTION

Arthropods play an important part in the transmission of human diseases; they also have a part in maintaining a reservoir of these diseases in nature. In many instances the only normal method of transmission is via an insect vector; in other cases insects act mainly as carriers or as protective agents of the pathogenic organisms or as culture media of these disease-producing microbes. In a recent list of the 44 principal protozoan and helminthic parasites of man at least 17 are transmitted to man only through the medium of arthropods (mainly insects); four others may be and frequently are transmitted through the agency of insects while several others may occasionally be so transmitted. In addition to these there are the spirochetal, rickettsial, bacterial and virus diseases of man; many, and in certain cases all, are transmitted by arthropods. At the present time the number of known human, animal and plant diseases that are transmitted by arthropods, mainly insects, is rather appalling. Furthermore each year sees new and, at times, startling developments in this field of inquiry.

It is interesting to recall that nearly all this knowledge has been acquired in less than half a century. The discovery by Fedtschenko (1869) that *Cyclops* species serve as the intermediate hosts of *Dracunculus medinensis* lay buried in Russian literature for many years. The work of Manson on *Wuchereria bancrofti* (1878-1883) had little influence on medical or scientific workers. The striking experimental and conclusive investigations of Smith and Kilbourne (1893) on Texas fever of cattle lay buried in a government report and apparently was unknown to most medical students. The discovery by Laveran in 1881 that malaria in man was caused by a parasite in the erythrocytes aroused great interest and much study was devoted to these parasites by Italian workers. Finally Ross, on the 20th of August, 1897, found the developing oocysts of the malaria parasite in the stomach of a mosquito which had bitten a patient four days previously. This eventful day was a milestone in

medicine and entomology. In celebration of the event Ross penned the following triumphant paean and sent it to his wife.

"This day relenting God
Hath placed within my hand
A wondrous thing; and God
Be praised. At his command,

Seeking His secret deeds
With tears and toiling breath
I find thy cunning seeds,
O million murdering Death.

I know this little thing
A myriad men will save.
O Death, where is thy sting?
Thy victory, O Grave?"

Long before Laveran and Ross had made their remarkable discoveries Carlos J. Finlay had propounded in 1881 his theory that *Aedes aegypti* (*Stegomyia fasciata*) was the vector of yellow fever. This theory he supported and largely demonstrated to be true by actual experimental work during the years 1882 to 1899. The confirmation of Finlay's work by the American Commission on Yellow Fever in 1900 and 1901 was brilliant and constitutes another great milestone in the history of medicine and entomology. The later remarkable developments in the study of insects as the vectors of pathogenic organisms of man, animals and plants are current history.

In dealing with insects as vectors of human diseases we may approach the subject either from the disease or the insect vector. As an entomologist I follow the latter course.

THE CLASS CRUSTACEA

The Crustacea play a small part in the transmission of human disease. Here *Cyclops* spp. (over 20 different species) serve as the infective host for man of the Guinea worm, *Dracunculus medinensis*. This parasite is widespread over central equatorial Africa, the northwest and west coasts of Africa, the Nile Valley, throughout the Near East, India; and, in the Western Hemisphere, in the Guianas, Bahia (Brazil) and parts of the West Indies. There is no doubt our soldiers can become infected

in those areas by drinking water contaminated with infected *Cyclops*. The parasite is also present in our own country in our fur-bearing animals but the ten human cases recorded from the United States and analyzed by Chitwood all appear to be imported cases.

The only other important parasite transmitted by Crustacea is the lung fluke, *Paragonimus westermanni*. This fluke is widely distributed in the Far East; it occurs also in New Guinea, the Cameroons, the Belgian Congo and Tripoli in Africa; and there are foci in Brazil, Peru and Venezuela. However, our troops are not likely to become infected unless they eat raw or partially cooked parasitized crabs or crayfish.

THE CLASS ARACHNIDA

The Order Acarina

The order Acarina is a large and difficult group with which the entomologist and medical man must concern themselves more and more Here we are, at present, concerned mainly with the superfamilies Ixodoidea, Parasitoidea and Trombidoidea. The Ixodoidea are of great importance.

The Ixodoidea or ticks are all external parasites of vertebrates The order contains two families: the Ixodidae or hard ticks and the Argasidae or the soft ticks. In the former the dorsal surface of the body is partially (females) or completely (males) covered by a shield or scutum; the head or capitulum projects in front and the males are usually smaller; in fact, much smaller than the females. Furthermore these ticks can take great quantities of blood and become of great size when fully engorged. In the Argasidae there is no dorsal shield or scutum, the capitulum is ventral and there is slight difference between males and females. The species feed frequently and never greatly increase in size when fully gorged.

The life cycles and feeding habits of ticks are rather ideal from the standpoint of disease transmission and dissemination. In the Argasidae the species feed to repletion in comparatively short intervals and undergo a considerable number of molts before reaching maturity. The adult males and females are comparatively long-lived and the females may deposit their eggs over a considerable period of time. Furthermore, they can withstand long periods, even several years, of starvation. In the Ixodidae the life cycles vary considerably. Normally the egg hatches and the larva must find a host. When replete the larva drops from the host, digests its blood meal and molts into the nymph. The nymph must now attach to a new host, usually the same host species, and when replete drops again, digests its blood meal and molts to the

adult stage. The adults, males and females, now attach to the same host species or may seek entirely new types of hosts. Mating usually takes place on the host. The males and females now drop from the host and the males normally die. The females after digesting their blood meals lay their eggs in very large masses, often as many as 6000 to 8000 eggs to a mass. The shrunken body of the female may usually be found near the egg mass. There are great variations in the life cycles of ixodid ticks and each species presents its own problem. However, we speak of one-host ticks (the entire life cycle, except egg-laying on one host); two-host ticks; and three-host ticks. In the Argasidae the various species are normally many-host ticks. Another point to be remembered is that all ticks can withstand considerable periods of starvation at each stage of their development while the adults of many species can survive several years without food.

Ticks and Disease

RELAPSING FEVER.—Relapsing fever or fevers are caused by species of *Spirochaeta*. No one has yet successfully defined species of spirochetes though more than twelve to fourteen have been named. Apparently no spirochete has been cultured on prepared media. Davis (1942) presents a more or less clear-cut method of distinguishing species of spirochetes based on his experimental data. Each species of *Ornithodoros* that is a relapsing fever vector carries a spirochete that is host-specific and this host-specific relationship offers a more accurate approach to the differentiation of spirochete species. Unfortunately Davis places too much reliance on the ability of taxonomists to diagnose what is a species in the genus *Ornithodoros*. We might revise the definition and state that any *Ornithodoros* tick that transmits a strain of *Spirochaeta* belongs to the same species as the original tick vector.

In tick-borne relapsing fever the spirochete must be obtained from some reservoir. In Central African relapsing fever, which is transmitted by *Ornithodoros moubata*, man and probably rodents are the reservoirs. *O. moubata* is, according to Buxton, restricted to East Africa from Egypt to Natal and extends to the west coast close to the mouth of the Congo but not elsewhere. This tick is primarily an inhabitant of man's abodes. It is found in his houses, mud huts, around camp sites, well heads, etc. Here the tick spends its entire life. On hatching from the egg the larva does not feed but molts to the nymphal stage. A succession of feedings and molts occur before the adult stage is reached. The species feeds primarily on man. The adult ticks are long-lived and the females lay their eggs over a considerable period of time.

The spirochete associated with this tick is *S. duttoni* and its developmental cycle in the tick is said to correspond to that which occurs in the louse (*Pediculus humanus*) except this spirochete invades the tissues and organs of the host, including the ovaries. The parasite invades the developing eggs and the young ticks hatching from the infected eggs can transmit the disease to new hosts. Transmission of the spirochete to man occurs during feeding but the exact method is somewhat in dispute. However, the spirochete gains entrance to the wound either from the salivary fluid or the coxal fluid or both, but not from the fecal wastes.

The epidemiology of Central African relapsing fever is characterized by its localization. In any community infected ticks may be restricted to special buildings or rest houses. Here the disease may occur regularly as the spirochete is hereditary in the tick and the ticks are not known to migrate. The disease is primarily a "house disease" and is not epidemic.

The question may be asked: Will this disease be introduced into America? If Davis' assumption is correct it should not, for it is extremely doubtful if the tick is introduced or, if introduced, it probably could not survive under our conditions except in the Southwest.

The other relapsing fevers that are tick-borne occur throughout the Mediterranean area but also extend southward from French Africa to Dakar in Senegal and eastward including most of the Arab, Turkish and Persian lands. These types of relapsing fevers do not occur in India or China. The tick vectors are *O. erraticus* in Tunis, Spain and Morocco; *O. tholozani* in the Near East (Persia, Arabia, Turkestan); and *O. verrucosus* in the Caucasus area and northern Persia. Whether these are distinct diseases is not known.

In the Western Hemisphere tick-borne relapsing fever is endemic in thirteen western states, British Columbia, Mexico, Central America (Guatemala), Panama, Columbia, Venezuela and Argentina. The known transmitters are *Ornithodoros parkeri*, *O. hermsi*, *O. talaje*, *O. rudis* and *O. turicata*. In all these ticks the spirochete can pass through the egg to the young so that it may be said to be hereditarily transmitted; the percentage of egg transmission is thought to be low.[1] The animal reservoirs of the spirochetes are rodents—chipmunks, tamarack squirrels and probably others. Davis believes that ticks are the natural reservoirs.

Rocky Mountain Spotted Fever.—This disease is now widespread in our own country. The etiological agent is *Dermacentroxenus rickettsi*, one of the rickettsias. The reservoir of the pathogen is in rodents,

[1] Davis (in a letter) informs me that with *O. turicata* he obtains close to 100 per cent transmission through the egg. See: Davis, Gordon. Relapsing fever. the tick *Ornithodoros turicata* as a spirochaetal reservoir. U. S. Pub. Health Repts. 58: 839–842. 1943.

particularly rabbits, jack rabbits, and squirrels. The tick vectors to man in our country are *Dermancentor andersoni*, *D. variabilis* and *D. occidentalis*. *D. andersoni* is the transmitter in the western mountain states; *D. occidentalis* in most of the Pacific coast states; *D. variabilis* throughout the rest of the country and parts of California and Oregon. The life cycles of these ticks make them rather ideal transmitters. It is not possible here to present in detail the life cycles and hosts of these ticks.

Dermacentor andersoni is a three-host tick. The fertilized female deposits her eggs, some 2000 to 8000, in some protected place on the ground. The period of oviposition requires about a month and takes place during the summer. The eggs hatch in from one to two months and the larvae attach to small mammals, especially rodents. Engorgement is completed in two to eight days and they drop to the ground, digest their blood meal and molt to the nymphal stage. Normally the species passes the first winter as unfed nymphs. The following spring the nymphs attach to the same hosts, rodents. When engorged the nymphs drop to the ground, digest their blood meal and molt to the adult. Normally the second winter is passed as unfed adults. The second spring the unfed adults now attach to the larger mammals as horses, cattle, sheep, mountain goat, deer, and man. Mating takes place on the host and when engorged they drop off. The males normally die and the females begin oviposition after digesting their blood meal. The normal life cycle occupies two years though there are many variations.

Dermacentor variabilis is also a three-host tick. Under favorable conditions the life cycle from egg to adult may be completed in less than two months or it may require nearly eight months depending on the time the eggs are laid, the available food supply and other factors. The larval and nymphal stages occur mostly on mice; mainly meadow mice, pine mice and white-footed mice. The adults prefer the larger mammals, particularly the dog. Man is readily attacked.

The ticks obtain the rickettsias (*D. rickettsi*) from their rodent hosts. These develop in the ticks and invade all tissues including the developing eggs. The infection passes from larva to nymph to adult and thence to the eggs. Man becomes infected when an infected tick engorges on him. In nature the pathogen is spread among the rodent hosts by infected larvae, nymphs and in the case of *D. variabilis* and *D. occidentalis* by the infected adults as they readily attach to many of the larger rodents. Again the rabbit tick, *Haemaphysalis leporis-palustris*, serves as a vector among rodents, especially rabbits. Parker and his associates consider

the rabbit tick one of the most important agents in the spread of the rickettsia among rabbits and thus maintains and extends the animal reservoir.

OTHER TICK-BORNE DISEASES.—The other tick-borne rickettsia diseases are exanthematous fever (fièvre boutonneuse), widely distributed about the Mediterranean Basin and transmitted by *Rhipicephalus sanguineus;* São Paulo fever of Brazil, transmitted by *Amblyomma cajennense*, *A. striatum* and *Rhipicephalus sanguineus;* "Q" fever of Montana and Australia, transmitted by *D. andersoni* in Montana; Tobia spotted fever of Colombia of which the vector is not known; tick-bite fever of east and south Africa of which the vectors are not certainly known though *Hyalomma aegyptium* and *R. sanguineus* are good experimental transmitters; Colorado tick fever transmitted by *D. andersoni;* and some other obscure diseases.

In addition tick paralysis is of some importance. *D. andersoni* is the criminal in America; *Ixodes pilosus* in S. Africa; *I. holocyclus* in Australia; *I. ricinus* in Europe; and certain other species in other parts of the world.

TULARAEMIA.—This bacterial disease (caused by *Bacterium tularense*) of man and animals is merely mentioned here. Ticks play an important part in the maintenance of the disease in nature and in human infections. The disease is widespread in this country, in Russia, Japan, Canada, various parts of Europe and probably other parts of the world. The wild reservoir is in rodents and also ground-loving birds. The reservoir is maintained by such ticks as *Dermacentor andersoni*, *D. variabilis* and *Haemaphysalis leporis-palustris;* by certain horseflies (Tabanidae) and some other parasitic insects. In ticks the infection is passed to the young in the eggs. The interrelations involved in the maintenance of the natural reservoir of this disease and human and animal infections are very complicated and time will not allow a full presentation here

The Other Acarina

Here should be mentioned the chiggers, *Trombidium irritans* (now designated as *T. alfreddugesi*) of the Americas, *T. autumnalis* of Europe, *T. hirsti* of Australia, and other species in various parts of the world. These extremely annoying mites are terrible pests to men in the army. In addition to the irritation and itchiness caused by them, their bites and the abrasion of the skin due to scratching are constant open wounds for more serious infections. Furthermore some of them are suspected to serve as vectors of various forms of typhus as "Scrub-Fever" in Malaya.

Trombidium akamuski is the vector of kedani or Japanese river fever. The mite is parasitic, in the larval stage, on mice, attaching largely in

and about the ears. The larvae feed for three or four days and then drop to the ground where they molt to the nymphal stage. The nymphal and adult stages are free living and said to feed on decaying organic matter or the juices of plants. The mice are the reservoir of the pathogen, *Rickettsia orientalis*. The rickettsia passes through the egg to the larvae and man becomes infected when larvae, carrying the parasite, feed on him. The disease is present in Japan, Formosa, China, Indo-China, the Malay Archipelago and Australia.

THE CLASS HEXAPODA

The Hexapoda or insects are familiar to everyone. Probably no group of animals is referred to more lightly by the average person, both educated and uneducated, than this vast and dominating class. They are present everywhere from the poles to the tropics, from the lowest valleys to the highest mountain tops and utilize the air more than any other group of animals. They have been captured floating in the air at more than 14,000 feet altitude and all evidence indicates a vast quantity of them are constantly floating about in air currents. Insects occupy practically every spot on the globe except the oceans, the inland seas and our deepest lakes. They constitute man's most important contenders for food and shelter, attack his person and rack his body with disease. Yet despite all this it is doubtful if man could survive without the many beneficial insects that do him untold services. We can discuss only a few important orders and families associated with the transmission of human diseases.

The Order Hemiptera

The species belonging in this order are mainly phytophagous. Only two families contain forms that attack man. The family Cimicidae, the bedbugs, and certain species of the assassin bugs, Reduviidae, are known to seek human blood. As far as known bedbugs are not directly responsible for the transmission of human disease. In the Reduviidae certain species of *Triatoma* (*Mestor*), *Rhodnius*, *Eratyrus*, and one or two other genera serve as the vectors of Chagas' disease. This disease, caused by *Trypanosoma cruzi*, occurs in parts of South America and Central America. The natural reservoir of the trypanosome occurs in rats, wood rats and certain other animals. Though the trypanosome occurs in our country, no human cases have thus far been reported.

The Order Anoplura

In this order are placed the biting and sucking lice. Here only the sucking lice are of primary interest. Though some 200 species of sucking

lice are known, only two species occur on man. These are *Pediculus humanus* var. *capitis* and var. *corporis* and *Phthirus pubis*. The latter species is not known to play any part in the dissemination of human disease.

The head louse, *Pediculus humanus* var. *capitis*, is present in varying numbers on man throughout the world. The abundance of head lice depends largely on the habits of the population. This is also true for the body louse, *P. humanus* var. *corporis*. Lice are usually more abundant when heavy clothing is worn and infrequently washed; or when people are crowded, ill-fed and poor; or where the hair of the head is allowed to grow with infrequent washings. In many tropical countries where the population is practically naked the body louse is not abundant though the head louse may be present in considerable numbers.

The life cycles of the head and body lice are very similar. The head louse lays its eggs normally on the hairs; the body louse prefers the clothing though it will lay them on the body hairs. Each female may lay 279 eggs at the rate of 9 eggs a day (Buxton). The female normally lives about 34 days. The eggs hatch in 9 days and the nymph becomes mature in 9 days. According to Buxton there is normally 30 per cent mortality of the eggs and 40 per cent of the nymphs. Allowing for this high death rate and starting with a single fertilized female, Buxton estimates a total population of nearly 15,000 eggs, nymphs and adults in 80 days. It will thus be seen the rapidity with which a louse population may develop under unsanitary conditions.

Lice and Disease

The principle diseases transmitted by lice are relapsing fever, "epidemic" typhus and trench fever.

Relapsing Fever.—The variety of this disease transmitted by lice is caused by *Spirochaeta recurrentis*. This louse spirochete cannot be transmitted by ticks (Buxton) and is said to be host-specific. The only reservoir of this spirochete must be man or the louse, for man is the only known host of the louse. The louse obtains the spirochete in taking blood from an infected individual. Within the louse, the fate of these spirochetes is differently interpreted by various students. Chung and Feng traced the spirochetes through the intestinal wall into the haemocele and within a few hours to several days the spirochetes multiplied, eventually being found in all the open spaces of the body and appendages. Here they persist as long as the louse lives. Unlike the tick spirochete (*Spirochaeta duttoni*) there is no transmission through the egg and no invasion of the tissues. Transmission takes place only by the crushing

or rupturing of infected lice on the injured skin or possibly through the uninjured mucous membranes. As lice travel readily from individual to individual, especially under crowded conditions, it is easy to see how the disease may spread.

Louse-borne relapsing fever is widespread in eastern and southern Europe, most of Asia, parts of North Africa and a vast belt stretching across Africa south of the Sahara desert east almost to the coast in Italian Somaliland. In South America it is recorded from Peru. It is also reported from Australia. It is doubtful if it occurs in North America though it seems to be indicated as present in northern California and perhaps Oregon.

Probably the most recent epidemic of louse-borne relapsing fever swept across Equatorial Africa south of the Sahara desert from Upper Guinea eastward, southward and northward. It is estimated that between 1921 and 1928 over more than 10 per cent of the population died from the disease.

Epidemic Typhus.—The etiological agent of typhus is *Rickettsia prowazeki*. The louse acquires the organism from typhus patients during the first ten days of the disease. The rickettsias invade the epithelial cells of the midgut where they multiply very rapidly, causing a great enlargement of the cells which finally rupture and discharge enormous numbers into the lumen. The rickettsias also multiply in the lumen. The breaking down of the epithelial cells brings about the death of the louse in about a week or ten days (Buxton). Human infection takes place through feces deposited on injured skin or via the conjunctiva or by crushing infected lice on scratches or abrasions. Furthermore the rickettsias can survive in the dry feces for more than two months, hence scratching the skin on which infected feces was deposited previously may bring about the disease. As lice are restricted to man the reservoir of the disease is not known, though the studies on "murine" or "endemic" typhus may solve this mystery.

Epidemic typhus is widespread in Europe, most of central and northern Asia, extensive areas of South America, Mexico, Central America and vast areas in Africa. It is said to be absent from most of India, China, the Malaysian area, Australia and North America.

Trench Fever.—Trench fever appeared as a distinct entity during the last war. As far as known it has completely disappeared. It will be interesting to see if it reappears during the present conflict. It is caused by *Rickettsia quintana* and is louse-borne.

The Order Diptera

The Diptera or flies are usually easily recognized as adults. They are generally small to minute, possess only a single pair of wings which are thin and membranous, and are active in the daytime or during twilight hours. There are many exceptions to such a generalization but the vast majority would be included. It is only among the blood-sucking and parasitic flies that we find the most marked exceptions. In their early stages as larvae or maggots we are confronted with the most extraordinary variety of habits and habitats. Furthermore, the recognition of species in both adults and larvae is extremely difficult and often almost impossible. The flies are undoubtedly the most important order of all insects in their relation to man and other animals as transmitters of disease-producing parasites, as scavengers, as troublesome pests and as agents in controlling many of the insects affecting agriculture. Here it will be only possible to present a brief summary of the most important groups.

THE FAMILY PSYCHODIDAE

The adult psychodids may be recognized by their hairy wings and their wing venation pattern. The subfamily Phlebotominae contains all the blood-sucking species. These are minute blood-sucking flies but unfortunately we know comparatively little about their biologies and, in many cases, less about the method of transmission of diseases. The flies are very small, pass readily through screens and bed nets unless the mesh is extremely fine. They feed only at twilight or during the night. Their breeding places are rarely found. A brief list of the known or suspected diseases associated with *Phlebotomus* is all that can be mentioned.

Pappataci or Three-day Fever. This is a disease of unknown etiology—a virus disease. It is widespread throughout the Mediterranean region, India, Ceylon and South China Its vector is *Phlebotomus papatasii*.

Kala Azar, Espundia and Oriental Sore.—These diseases are caused by species of *Leishmania* and are widely distributed in the regions where active military operations are in progress. How they are transmitted remains somewhat doubtful but all the evidence indicates that *Phlebotomus* species are involved.

Verruga Peruviana or Oroya Fever.—This may play no part in the war as it occurs only in certain mountainous areas of Peru and is reported from certain small areas in Ecuador. However, the mineral and other resources of those regions are important. Hertig (1942)

reaches the conclusion that only *Phlebotomus verrucarum* is involved in the transmission of the disease.

THE FAMILY CULICIDAE

The mosquitoes constitute the most important group involved in the transmission of human and other animal diseases. Furthermore the blood-sucking habits of most of the species cause great distress and make life almost unendurable in certain areas of the world. Our troops suffer untold agonies from their bites and the difficulties of protecting troops in the field, on sentry duty and in similar work are not easily solved.

Adult mosquitoes are easily recognized as they are practically the only flies in which the wing veins and margins are furnished with delicate scales. One subfamily, the Chaoborinae, contains no blood-sucking forms and the mouthparts are relatively short and not adapted for piercing. The subfamily Culicinae contains all the blood-sucking species. This subfamily is divided into three tribes of which the most important are the Anophelini and Culicini. The total number of world species is about 1700. Of these about 200 are anophelines and the culicines include most of the remainder.

All species of mosquitoes breed in water and their characteristic larvae are easily recognized by anyone who will study them in the most superficial manner. In this presentation it is assumed that the general outline of mosquito biology is fairly well known and we can restrict these brief remarks to certain special features.

The Anophelines

Adult anophelines can be recognized by their resting habits and they generally have spotted wings. In resting, anophelines hold the body at an angle of 30° to almost 90° to the support; in culicines the body is held nearly parallel with the head only bent down. The most distinctive characters of anophelines are the crescent-shaped scutellum and a long straight proboscis with the palpi as long as the proboscis. No other mosquitoes have these associated characteristics. The larvae lack an anal siphon, normally rest and feed at the water surface and parallel to it and are provided with peculiar hair tufts known as "float hairs" or "palmate hairs" on a number of the abdominal segments. The pupae can be distinguished but not easily. The eggs are laid on the water surface and are provided with characteristic floats.

At the present time nearly 200 species of anophelines are known. These are placed in three genera,—*Chagasia* (3 species), *Bironella* (6

species) and *Anopheles* (all the remaining species). It is not possible here to present an account of anopheline biology. We have learned that species biology is the key to the study and control of anophelines and the associated diseases they transmit.

The Culicines

Adult culicines are all the other mosquitoes that lack the combinations of characters indicated for anophelines. The larvae possess an anal siphon, feed and rest hanging at a distinct angle to the water surface. The eggs are laid in various places but always on the water, near water or where water may be expected to occur. The eggs never possess floats. The culicines are divided among some 28 to 30 genera. It is not feasible to present a general summary of the biology of the culicines as species biology is our only key to effective means of controlling them and the associated diseases they transmit.

Mosquitoes and Human Disease

MALARIA.—Malaria is probably the greatest scourge of mankind. This statement seems to be adequately supported by what is occurring in the great battlefields of the present war. If you look at a world distribution map of highly endemic areas of malaria it may be seen that great masses of troops are being concentrated there. Man and anophelines are the sole propagators and reservoirs of malaria. In order that malaria may occur in any region there must be a human reservoir of viable gametocytes of one or all the species of human *Plasmodium*. Then there must be present species of anophelines that seek human blood and in which the sexual and sporogonous cycle of the parasites can be completed successfully. Our knowledge of anophelines indicates that many anophelines do not seek human blood and hence do not play any significant part in the dissemination of malaria. However, experimental data indicate that practically any species of *Anopheles* can be infected and the gametogonous and sporogonous cycles completed. However, if such species do not or only rarely feed on man they can only play a minor role in malaria transmission.

At the present time only some 20 to 30 species of anophelines seem to be incriminated as dangerous transmitters of malaria. By concentrating on a study of the bionomics of these species and adapting our control measures to species sanitation, it is always possible to rid any region of malaria or reduce it to minor importance if funds, labor and adequate scientific direction are available. Such work has been accomplished in certain parts of the world where endemic and epidemic malaria are

hazards to great governmental undertakings or vast business enterprises. Today our Army and Navy are face to face with a vast malaria problem and it is difficult to see how it can be successfully met under battle conditions.

What are the principal malarial transmitters throughout the world? We do not know them all and furthermore we do not know the bionomics of many species with which our armed forces must deal. If we did, plans could be made in advance and carried out very efficiently even under war conditions (see TABLE 1).

TABLE 1

THE MORE IMPORTANT MALARIA-TRANSMITTING ANOPHELINES OF THE WORLD

Species	Breeding places	Distribution
	A. NEARTIC REGION	
A. freeborni Aitken	Fresh clear seepage water, rice fields, irrigation ditches and similar places	Interior valleys New Mexico, California, and Oregon
A. quadrimaculatus Say	Fresh-water pools, ponds, margins of lakes, swamps, cat-tail marshes, and similar places	Eastern United States from New Hampshire to southern Ontario, west to Minnesota and south to central Texas
A. pseudopunctipennis Theobald	Clear water rich in algae and exposed to sunlight	Mississippi Valley north to central Oklahoma, Texas, Mexico, Central America, Grenada, Trinidad, South America except Brazil
	B. NEOTROPICAL REGION	
A. albimanus Wiedemann	Sunlit exposed collections of water, stagnant or pure, fresh or brackish; lake breeder when aquatic vegetation covers water	Southeast Texas, Mexico, Central America, northern South America, West Indies quite generally
A. aquasalis Curry (*tarsimaculatus*)	Brackish-water areas and at times in fresh water (rice fields)	Central America, Venezuela, Guianas, Brazil, West Indies
A. bellator Dyar & Knab	Breeds in water at leaf bases of bromeliads	Trinidad, Venezuela, Brazil
A. darlingi Root	Pools with vegetation; in surface mats of vegetation; shaded pools and lagoons	Central America, northern South America, Brazil, Argentine
A. gambiae Giles (*costalis*)	Pools, slow streams, hoof-prints, puddles, seepage water, etc.	Region of Natal, Brazil (now thought exterminated there)
A. pseudopunctipennis Theobald	Clear water rich in algae and exposed to sunlight	Mississippi Valley north to central Oklahoma, Texas, Mexico, Central America, Grenada, Trinidad, South America except Brazil

TABLE 1 (*Continued*)

Species	Breeding places	Distribution
	C. PALEARCTIC REGION	
A. algeriensis Theobald	Stagnant weedy water, clear, slow-flowing streams with aquatic plants	Mediterranean region east to Mesopotamia and Turkestan
A. hyrcanus var. *sinensis* Wiedemann	Open swamps, pools, rice fields, drains, canals, and slow-flowing streams	North and Central China, Korea, Japan; Mediterranean region, Mesopotamia
A. maculipennis (varieties) Meigen	Great variety of waters both fresh and brackish. Details of varieties not complete	Europe, North Africa, Central and part of northern Asia. Wide distribution
A. multicolor Cambouliu	Breeds in saline desert waters (salt pans), small pools	North Africa, Egypt, desert oases, west to Baluchistan
A. pharoensis Theobald	Rice fields, flooded lands and grassy pools.	Egypt, Palestine
A. sacharovi Favr (*elutus* Edwards)	Stagnant pools not foul, puddles, margins of slow-flowing streams	Southern Europe, North Africa, Asia Minor, Palestine
A. superpictus Grassi	Pools in beds of hill streams or mountain rivers in irrigation channels	Spain, Italy, Balkans, Greece, Algeria, eastern Mediterranean area, Mesopotamia, northwestern India, Turkestan
	D. ORIENTAL REGION	
A. aconitus Donitz and var. *filipinae* Manalang	Swamps, ponds, pools with grassy edges, rice fields, irrigation ditches	India, Ceylon, Burma, Siam, Borneo, northern East Indies, Philippines; wide distribution
A. annularis Van der Wulp	Weedy stagnant waters, margins of lakes, tanks, moats, rice fields, drains, stream pools, swamps	India, Ceylon, Burma, Malayan area, Borneo, Philippines
A. barbirostris Van der Wulp	Deep stagnant water with much vegetation and preferably in shade as margins of lakes, swamps, sluggish rivers, rice fields	India, Ceylon, Burma, Malayan region, New Guinea
A. culicifacies Giles	Clean fresh water, irrigation channels, pools in canals, rivers, slow-moving streams, temporary rain pools; also in brackish water	India, Ceylon, Siam, China (Yunnan)
A. fluviatilis James (*listoni* Giles)	Pools in stream beds, and slow-flowing streams with vegetation, springs, edges of swamps, ponds, rice fields	India, Ceylon, Siam, Tonkin, Turkestan
A. leucosphyrus Dönitz	Breeds in deep jungle and forest pools, beside rocky streams, shaded swamps	India, Ceylon, Burma, Siam, French Indo-China, Malay Peninsula, Netherland East Indies, Borneo

TABLE 1 (*Continued*)

Species	Breeding places	Distribution
A. maculatus Theobald	Stream and river-bed breeder, fast-flowing streams with grassy edges and small pools	India, Ceylon, Burma, Malay Peninsula, Netherlands East Indies, Borneo; French Indo-China, Siam, Philippines
A. minimus Theobald and var. *flavirostris* Ludlow	Slow-running streams with grassy edges; margins of swamps, ditches, ponds	India, Ceylon, Burma, Assam, Indo-China, Siam, southern China, Formosa, Malay Peninsula, Netherlands East Indies, Philippines
A. subpictus Grassi	Small pools, excavations, burrow pits, artificial containers, roof gutters, rice fields, irrigation ditches	India, Ceylon, Burma, Netherlands East Indies, Malayan region, Celebes, Moluccas, Philippines, southern China, French Indo-China
A. sundaicus Rodenwaldt	Brackish-water pools	Eastern India, Burma, Andaman Islands, Malayan region
A. umbrosus Theobald	A jungle breeder in stagnant pools, pockets of water among trees, mangrove swamps and shaded mountain brooks	Malayan region, Celebes, eastern India
A. stephensi Giles	Pools in rivers and stream beds, sluggish creeks, drains, wells; also in brackish water	India, Burma, lower Mesopotamia
E. AUSTRALIAN REGION		
A. annulipes Walker	Pools, marshes, creeks, brackish water?	Australia, Tasmania, ?New Hebrides
A. punctulatus Dönitz	Pools, swamps, puddles, stagnant water	Northern Australia, New Guinea, Solomons, New Hebrides, New Britain, ?New Caledonia
A. punctulatus var. *moluccensis* Swellengrebel and Sw. de Graaf	All kinds of water, fresh or brackish, artificial collections in coconut shells, containers, etc.	New Britain, New Guinea, Moluccas, Solomons
F. ETHIOPIAN REGION		
A. funestus Giles	Swamps, weedy margins of streams, rivers, furrows, ditches, ponds, seepage areas	Tropical Africa south to Natal; widely distributed; Mauritius
A. gambiae Giles (*costalis*)	Open pools, hoofprints, puddles, seepage, water holes, drains, pools in stream beds	Practically all of Africa south of Sahara desert; Madagascar, Mauritius
A. hargreavesi Evans	Open swamps, among *Pistia*, in clear water in open jungles, grassy borders of streams	West Africa, southern Nigeria, Belgian Congo

TABLE 1 (*Continued*)

	Breeding places	Distribution
A. moucheti Evans	Grassy margins of streams, rivers, pools and ponds with vegetation	Central and eastern Belgian Congo, Uganda, ?Tanganyika
A. nili Theobald	Along shaded banks of clear flowing streams and rivers, occasionally in swamps and ditches	Across tropical Africa, local in its occurrence

The problem of malaria among our troops will be a serious one. Another important phase is the transportation of "good" malaria transmitters from one region to new areas and the introduction of malaria to nonimmune populations. The results of such introductions are well exemplified by the terrible epidemic in Mauritius (1865 1867) in which more than 25 per cent of the population perished in one year and the most recent ones in Brazil (1931 and 1938) where in certain areas 90 per cent of the population were infected and an estimated death rate of 10 per cent. The movements of ships and planes today means that the islands of the Pacific may all become the home of malaria-transmitting mosquitoes, even including the Hawaiian Islands. If our Japanese enemies determined to hit us hard they could introduce a good malaria transmitter to the Hawaiian Islands and probably the result would be catastrophic. It will be difficult to avoid the spread of mosquitoes as they readily invade planes and ships. On planes, I am informed, the recesses where the landing gear is withdrawn is an ideal place for mosquitoes. Will such places be ideal for transporting mosquitoes and some other insects?

DENGUE.—Dengue is widespread throughout a vast area of the world. Our troops are largely concentrated in dengue areas as in North Africa, the islands of the Pacific and Australia. Dengue is transmitted by only two species of mosquitoes, *Aedes aegypti* and *A. albopictus*. It is a virus disease. The mosquito can be infected by taking blood during the first three days of the febrile attack. About eleven days are required for the developmental cycle in the mosquito. Once infected the mosquito is capable of transmitting the disease as long as it lives, in some cases nearly three months.

The disease is not serious but is debilitating and often appears in epidemic form. In the Athens epidemic (from Sept. 1927 to Sept. 1928) some 90 per cent of the population suffered from the disease and in Athens alone some 239,000 cases were reported. In Miami in July,

1934, over 1000 cases were reported and within a month some 6000 cases developed. The epidemic spread to Jacksonville and thence to Georgia where several thousand cases were reported. This sudden epidemic, like the one in 1922 in Texas, illustrates the prevalence of *Aedes aegypti* in our South.

YELLOW FEVER.—Yellow fever may not play a serious role in our armed forces. An excellent vaccine is available and this should provide adequate protection. However, the widespread distribution of the disease in Africa and South America indicates a potential danger to the native population under wartime conditions. Furthermore, the spread of the disease to the Oriental and Australian regions is always possible and probable. In South America there is an extensive reservoir of jungle yellow fever and probably also in Africa. The mosquito vectors are present throughout nearly all regions of the world but most predominant in the tropical and subtropical areas. The reservoir of the disease is recorded as primarily among animals. The following are listed by the Rockefeller Foundation report for 1940:

Primates	Man and monkeys
Marsupials	All species of opossums
Edentates	Anteaters, sloths and armadillos
Rodents	Agouti, paca, capybara, some species of mice

In addition we should not forget that *Aedes aegypti* retains infection for a long period; Bauer (1940) records keeping an infected individual alive for 200 days. How long this mosquito can live in the wild and retain the ability to transmit the disease is not positively known (see TABLE 2).

FILARIASIS.—Filariasis due to *Wuchereria bancrofti* (TABLE 3, A) is prevalent in nearly all the areas where our troops are at present fighting or where they are located in large numbers, except in Great Britain. Though invasion by filarial worms may not affect our troops for some time, yet those infected will be an excellent source to infect our native mosquitoes. Some 7 *Aedes*, 28 *Anopheles*, 6 *Culex* and 4 *Mansonia* species are known to be efficient transmitters. Of these 45 species 20 have been found infected in the wild.

In addition we have *Filaria malayi* (TABLE 3, B) distributed throughout the Netherlands Indies, Borneo, New Guinea, Travancore (India) and about Huchow in China. The intermediate hosts are mosquitoes. *Onchocerca volvulus* (TABLE 3, C) is distributed widely in tropical Africa, southern Mexico and parts of Guatemala. The intermediate hosts are species of black flies (Simuliidae).

TABLE 2
MOSQUITOES INVOLVED IN THE TRANSMISSION OF YELLOW FEVER

Species	Breeding places	Distribution
Aedes aegypti[a]	Artificial containers	Tropics and subtropics
Aedes africanus[c, d]	Tree holes, stumps, artificial containers	Ethiopian
Aedes albopictus[c, d]	Artificial containers	Oriental and eastern Nearctic regions
Aedes fluviatilis[c, d]	Rock holes; clay rings, ant rings	Neotropical
Aedes fulvithorax[c, d]	Tree holes	Neotropical
Aedes geniculatus[c, d]	Tree holes	Palearctic
Aedes irritans[c, d]	Crab holes, salt areas inland	West Africa
Aedes leucocelanus[b, d]	Tree holes	Neotropical
Aedes nigricephalus[c, d]	Crab holes	Ethiopian (West)
Aedes nubilus[c, d]	Temporary ground pools	Neotropical
Aedes punctocostalis[d]	Crab holes	Ethiopian (West)
Aedes simpsoni[b, d]	Tree holes, bamboo stems; leaf axils	Ethiopian
Aedes serratus[d]	Temporary rain pools	Neotropical
Aedes stokesi[c, d]	Tree holes, banana and bamboo stumps	Ethiopian
Aedes terrens[d]	Tree holes	Neotropical
Aedes triseriatus[c, d]	Tree holes	Nearctic
Aedes scapularis[c, d]	Rain pools	Neotropical
Aedes taeniorhynchus[d]	Saline marshes	Nearctic, Neotropical
Aedes vittatus[c, d]	Rock pools, drains, artificial containers	Ethiopian, Oriental
Eretmopodites chrysogaster[c, d]	Cacao husks, coconut shells, snail shells, artificial containers	Ethiopian
Culex fatigans[c, d]	Domestic; all sorts of containers	Tropical and semitropical
Culex thalassius[c, d]	Crab holes, inland salt areas	Ethiopian
Taeniorhynchus africanus[c, d]	Larvae and pupae attached to aquatic plants	Ethiopian
T. uniformis[c, d]	As above	Ethiopian, Oriental Australian
T. albicosta[d]	As above	Neotropical
T. fasciolatus[d]	As above?	Neotropical
T. amazonensis[d]	As above?	Neotropical
T. titillans[d]	Larvae and pupae attached to aquatic plants	Neotropical
Psorophora cingulata[d]	Temporary rain pools	Neotropical
P. ferox[d]	Temporary rain pools	Nearctic and Neotropical
Haemagogus capricornii[b, d]	Water in tree holes	Neotropical

[a] Principal vector.
[b] Known infected in the wild.
[c] Experimental transmission.
[d] Good incubators.

TABLE 3

INSECT TRANSMITTERS OF FILARIASIS

Species	Infection reported	General distribution
	A. Wuchereria bancrofti	
Aedes aegypti[a]	West Africa; New South Wales	Tropical and subtropical regions
Aedes albolineatus[a]	Malaya	Malayan region; Solomons
Aedes caspius[b]	Palestine	Europe, North Africa, Asia Minor, Central Asia, Punjab
Aedes desmotes[b]	Malaya	India, Malaya, Philippines
Aedes mediopunctatus[b]	Malaya	Malaya, India, Ceylon
Aedes scutellaris[b]	Fiji, Pacific Islands	Papua, Solomons, Amboina, Philippines
Aedes togoi[a]	Japan	Japan, China, eastern Siberia
Anopheles albimanus[b]	Caribbean area	Florida, southeastern Texas, Central America, Venezuela to Brazil, West Indies
Anopheles albitarsis[b]	Brazil	Argentina, Brazil
Anopheles algeriensis[b]	North Africa	Mediterranean region
Anopheles amictus[b]	North Queensland	Tropical Australia
Anopheles bancrofti[a]	New Guinea	Northern Australia, New Guinea, Philippines, Ceylon
Anopheles barbirostris[b]	India, Celebes	India, Ceylon, Malayan region, Philippines
Anopheles gambiae[a]	Africa	Africa, Madagascar
Anopheles coustani ?[b]	Mauritius	
Anopheles annularis[b]	India	India, Malayan region, Philippines
Anopheles funestus[a]	Africa	Tropical Africa, Mauritius
Anopheles hyrcanus var. *nigerrimus*[a]	Travancore	India, Ceylon, Malaya, Borneo, Philippines
A. h. var. *sinensis*[a]	Shanghai, Siam	China, Japan, Formosa
A. jeyporiensis[a]	Hong Kong	Eastern India, Formosa
A. ludlowi[a]	India	Eastern India, Malayan region, Philippines
A. maculatus[a]	Hong Kong	Oriental region
A. maculipalpis[b]	Mauritius	Africa, Mauritius
A. mauritianus[a]	Mauritius	Africa, Madagascar
A. minimus[a]	Hong Kong	Eastern India, Malaya, Hong Kong, Philippines
A. pallidus[b]	India	India
A. philippinensis[a]	India	Burma, Malaya, Philippines
A. punctulatus var. *moluccensis*[a]	New Guinea	New Guinea, Moluccas, New Hebrides, New Caledonia, Solomons, northern Australia
A. rhodesiensis[b]	Sierra Leone	Tropical Africa
A. splendidus[a]	Hong Kong	India, southern China, Formosa
A. squamosus[a]	Sierra Leone	Africa, Madagascar
A. subpictus[b,i]	India	Oriental region

TABLE 8 (*Continued*)

Species	Infection reported	General distribution
A. stephensi[b]	India	India, Lower Mesopotamia
A. subpictus[b]	India	India, Malayan region
Culex fatigans[a]	Widespread	Tropical and subtropical regions
Culex fuscanus[b]	China	Oriental region
Culex gelidus[b]	Malaya	Oriental region
Culex pipiens[a]	China, Japan, Egypt	Temperate regions
Culex sitiens[b]	Malaya	Coastal regions from east Africa to Fiji
Culex whitmorei[b]	Malaya	Oriental region
Taeniorhynchus africanus[b]	Africa	Ethiopian region
Taeniorhynchus indianus[a]	Tonkin	Burma, Siam, Indo-China, Java
Taeniorhynchus juxtamansonius[b]	Brazil	Brazil
Taeniorhynchus uniformis[b]	Africa	Africa, Oriental region, northern Australia

B. *Filaria malayi*

Taeniorhynchus annulatus[a]		Oriental region
Taeniorhynchus annulifera[a]		Oriental region
Taeniorhynchus bonneae[a]		Oriental region
Taeniorhynchus uniformis[a]		Africa, Oriental region
Taeniorhynchus indianus[a]		Burma, Siam, Indo-China, Java
Anopheles barbirostris[a]		Oriental region
Anopheles hyrcanus var. *sinensis*[a]		Oriental region
Anopheles punctulatus[a]		Australiasian region
Anopheles minimus[a]	Indo-China	Oriental region
Anopheles jeyporiensis[a]	Indo-China	Oriental region
Culex fatigans[a]		Tropical and subtropical
Culex pallidothorax[a]		Oriental region

C. *Onchocerca volvulus*

Simulium damnosum; S. aridum; S. ochraceum; S. neavei; S. mooseri (all black flies; family, Simuliidae)

[a] Natural infection.
[b] Experimental infection

HUMAN ENCEPHALITIS.—Equine encephalomyelitis, a virus disease, was recognized in horses in the U. S. before 1900 but the virus was not isolated till 1931. Since then two strains, an eastern and a western, have been isolated. Equine encephalomyelitis is widespread in Canada, parts of South America, Panama, Central Europe, Russia, India, and Japan.

Human infections were first recorded in Massachusetts in 1938 and

California the same year. Since then the disease in humans has been rather rare till the great outbreak in the Northern Plains States and Prairie Provinces of Canada in 1941. In that year more than 3000 human cases are recorded with a death rate varying from 9 per cent to 16 per cent.

Mosquitoes have been incriminated and some 12 species of mosquitoes have been shown capable of transmitting the disease to experimental animals. In 1941 *Culex tarsalis* was found naturally infected in the state of Washington. Also the tick, *Dermacentor andersoni*, has been shown to transmit the disease experimentally.

The reservoir of the disease is recorded in a variety of animals including birds and the domestic fowl.

THE FAMILY MUSCIDAE

The family Muscidae is difficult to define or indicate the limits of the forms included. It is composed of at least two subfamilies, the Stomoxydinae, and the Muscinae. The former includes all the well-known blood-sucking species; the latter, all the more common flies about our homes.

The blood-sucking forms belong mostly to the genera *Stomoxys*, *Haemotobia*, *Stygeromyia*, *Haematobosca*, *Bdellolarynx*, *Glossina*, and a few others.

Stomoxys calcitrans.—This biting stable fly is widely distributed throughout the world. It is a vicious blood-sucker and is most active in bright sunny weather. Though, undoubtedly, it prefers animal blood, yet it readily attacks man and when abundant renders life almost unbearable in the open. It is not known to serve as a vector of disease except in a mechanical manner through interrupted feedings. In this manner it can transmit such diseases as anthrax, tetanus, trypanosomiasis, and probably other blood infections.

THE GLOSSINA OR TSETSE FLIES.—The tsetse flies are restricted to Africa. They occur in the tropical and subtropical regions extending south of a line drawn from the mouth of the Senegal River on the west coast east through Lake Chad and Lake Rudolph to a point about 4° N. Latitude on the east coast; the southern line extends from the mouth of the Cunene River east along the southern border of Angola, thence southeasterly to Zululand. Within this area there are some twenty species of *Glossina* flies and a few subspecies. Though they occur widespread in this area, the various species are restricted to more or less local habitats usually called "fly belts."

The adults are long lived; some species live more than 250 days.

They can migrate considerable distances but due to their special requirements for food, shelter and reproduction they occupy particular areas ("fly belts"). Both males and females are vicious blood-suckers. The females do not lay eggs, but a single egg at a time passes into a uterine cavity where it hatches and the larva obtains its food from the so-called "milk-gland." The larva matures in from 10 to 12 days. The mature maggot is then laid in dry soil, always in the shade, and, in some species, in close proximity to water. The larva pupates and the pupal period varies from 21 to more than 60 days, depending on the species and the temperature. The reproductive capacity of the various species is not well known.

Glossina Flies and Disease

Glossina flies are the intermediate hosts and vectors of many African trypanosomes. They are the only known vectors of human sleeping sickness caused by *Trypanosoma gambiense* and *T. rhodesiense*. Gambian sleeping sickness, caused by *T. gambiense*, now extends across Africa from about 15° N. to 15° S. Latitude In this area there are many regions where the infection does not exist, though other regions in this area show a high rate of infection. Rhodesian sleeping sickness, caused by *T. rhodesiense*, is restricted to Rhodesia (northeast and south), Nyasaland, Portuguese East Africa, Tanganyika and parts of Mozambique.

Glossina palpalis is the principal vector of *T. gambiense*, *G. morsitans* is the important vector of *T. rhodesiense*. These flies obtain the trypanosome from persons suffering from the disease. Within the fly the trypanosomes undergo a cyclic development in the fly's intestine, and eventually pass up the esophagus and gain entrance to the salivary glands. Here they develop to the infective stage. This cyclic development requires about twenty days. The infected fly transmits the trypanosomes while feeding, and can transmit them as long as it lives.

The other species associated with the transmission of *T. gambiense* are *G. morsitans*, *G. fusca*, *G. pallidipes* and *G. tachinoides*. In the case of *T. rhodesiense*, *Glossina swynnertoni* is also an important vector.

The Order Siphonatera

Fleas, as adults, are all intermittent external parasites of warm-blooded animals. Man is attacked by a number of species and certain species serve as vectors of *Bacillus pestis* (plague) among rodents and to man, others are intermediate hosts of some helminths.

Tunga penetrans, the chigoe, is widely distributed in tropical America

and Africa. The habit of the female in burying herself in the skin of the host makes it a serious pest of man. In man the favorite places of attack are between the toes and under the toe nails. The enlarging females cause intense itching and inflammation. Ulceration and secondary infections commonly occur. This may result in gangrene and not infrequently tetanus.

Fleas as vectors of *Bacillus pestis*, the causative agent of plague, may play a serious role in this war. Plague is widespread throughout most of the areas where great concentrations of our troops are at present, and human cases have been reported in these areas within the past ten years. Furthermore, sylvatic plague is now present in many of these areas. With our modern sanitary conditions it is hoped plague may not occur among our troops. However, the stress of war, famine and other hardships may bring about an epidemic of plague in the native populations and this may prove serious. The more important species of fleas concerned in the maintenance of plague among rodents are:

Oropsylla montana—Among squirrels. Western United States and Mexico.

Hoplopsyllus anomalus—Spermophiles. Western United States.

Oropsylla silantjewi—Marmots. Mongolia.

Ceratophyllus tesquorum—Ground squirrels. Asiatic Russia.

Leptopsylla segnis.—Mice, rats. Europe, North America.

Rhopallopsyllus cavicola—The cavy. Argentine and Ecuador.

Xenopsylla eridos—Gerbilles. South Africa.

Xenopsylla brasiliensis—Rodents. Africa, India, South America.

Nosopsyllus fasciatus—Rats. Cosmopolitan in temperate regions.

Xenopsylla astia—Rats. Oriental region.

Xenopsylla cheopis.—Rats. Widespread in many parts of the world.

The species of fleas that are important in transmitting bubonic plague from rats to man are:

Xenopsylla cheopis
Nosopsyllus fasciatus
Xenopsylla astia.
Xenopsylla brasiliensis

The human flea, *Pulex irritans*, may transmit plague among humans during an epidemic but is considered of minor importance. In this same category are the cat and dog fleas (*Ctenocephalides felis* and *C. canis*).

Fleas and Endemic Typhus.—In recent years, fleas have been shown to serve as transmitters of endemic or murine typhus. This is said to be a mild form of "epidemic" or "Old World" typhus. Its

reservoir is in mice and rats. Its principal vectors to man are the fleas, *Nosopsyllus fasciatus*, *Xenopsylla cheopis*, and probably others.

DISCUSSION OF THE PAPER

Dr. C. H. Curran (*The American Museum of Natural History, New York*):

This subject has been covered so thoroughly by Dr. Matheson that there is nothing that can be added. However, I should like to make a few remarks upon some phases of diseases carried by insects which might lead to further discussion and possibly to new research. Although we have learned a good deal in the short time that we have known of the role of insects as disease carriers, we may safely say that we still know almost nothing about them. We need more men engaged in this work—many more.

I have long been interested in typhus and trench fever, diseases carried by lice, because I had contact with the latter in the last war. There is very good evidence to indicate that when typhus is transferred by fleas from rat to man it is of a mild form and produces immunity, but when transmitted by lice from man to man it becomes very virulent and the death rate may be high. I think it possible that a similar situation may exist in the case of sylvatic plague, infesting rodents in the United States. Since plague is so widespread in this country and no outbreaks have occurred among hunters and trappers, despite exceedingly large numbers of bites by fleas, it is apparent that there is some limitation on the transmission of sylvatic plague from rodents to man. Indeed, we cannot overlook the possibility that man is not affected by this rodent form of plague.

There is one other disease which I should like to mention—infantile paralysis. I have no doubt that this disease is carried by insects and I also believe that the stable fly, *Stomoxys calcitrans*, is the chief vector. Brues and others have demonstrated that the disease can be transmitted by this fly. The transmission must be purely mechanical. It would seem, therefore, that the disease can be carried only as a result of interrupted feeding. The fly must have the opportunity to suck some blood from a victim of the disease, but not to obtain a satisfying meal. If hungry it will bite again in a very short time and it could then transmit the disease. If, on the other hand, the fly obtains a full meal it will not feed again for several days. The case against the stable fly is very strong, but it seems certain that other biting insects are also vectors. However, *Stomoxys* is most abundant at the time that the disease is at its height and since the disease is a "rural" one, and the fly is associated with horses and cattle and abundant about bathing beaches, it seems to be the logical carrier.

Nothing is known of the natural reservoir of poliomyelitis. The discovery of this would be a long step in fighting the disease. I should like to suggest the possibility that there might be some connection between "distemper" in dogs and infantile paralysis. Some forms of distemper resemble poliomyelitis closely. Moreover, both diseases are most prevalent at the same season of the year, the so-called "dog-days." I realize, of course, that there is no real evidence to support this theory, but at the same time I feel that it is worthy of investigation

Dr. N. R. Stoll (*Rockefeller Institute for Medical Research, Princeton, N. J.*):

It is noteworthy that filariasis, the disease omitted in the formal paper, has caused more discussion than the ones mentioned. At the city of Soochow, China, I have noted that a mild elephantiasis was quite localized just outside of one of the city's gates almost to the exclusion of other areas.

A mosquito takes microfilaria to the skin but does not inject it. The *Filaria* probably enter through the wound. Infected mosquitoes are less viable than mosquitoes infected with malarial parasites.

Dr. G. C. Shattuck (*Harvard Medical School, Cambridge, Mass.*):

We should not lose sight of the fact that bacterial virus and spirochetal diseases are important, just as are animal parasites.

Dr. L. T. Coggeshall (*University of Michigan, Ann Arbor, Mich.*):

There are also possible relations between climatic cycles and fatal outbreaks of malaria. One such outbreak in Ceylon appeared to be the result of a freak climatic cycle.

Dr. D. L. Augustine (*Harvard Medical School, Cambridge, Mass.*):

How many vectors have been removed from planes?

Reply by Dr. Matheson:

There are not many cases of definite identification of insects brought by planes. Dr. Coggeshall has shown that most of the insects thus arriving were dead. (Customs of current fumigation practice of airplanes were reviewed at this point.)

The real hazard is not from usages of the established air lines, but from military planes operating under the special conditions incident to war emergencies.

Insects known to have arrived in America on planes include one tsetse fly and several *Anopheles gambiae*.

Dr. H. Fox (*New York University, New York, N. Y.*):

It should be noted that *Leishmania* can be transferred not only by vectors but also from person to person.

CLINICAL FEATURES OF PARASITIC DISEASES AND THEIR CONSIDERATION IN MILITARY AND NAVAL OPERATIONS

By

THOMAS T. MACKIE

Lieutenant Colonel, Medical Corps, Army Medical School, Washington, D. C.

It is apparent that the present war is to be fought, to a great extent, in the tropics and in areas adjacent to the tropics in which many tropical and parasitic diseases are widely endemic. In consequence, the great numbers of men who necessarily will participate in the overseas operations of our expeditionary forces are nonimmunes entering hyperendemic areas. This immediately presents the complicated problem of protection adequate to prevent a high disease morbidity rate and consequent serious interference with the conduct of military operations

In the last war the experience of the armies in Gallipoli, Macedonia, the Near East, and East Africa pointed to the gravity of the problem and the hazard which tropical disease may create. Because of the inroads of malaria, the French army in Macedonia in 1916 could put into the field only 30,000 men out of a force of 120,000. The British in the same area had 30,000 cases of malaria among their troops in 1916, and 70,000 in 1917. The Germans were likewise similarly affected, so that *Plasmodium* alone was responsible for the immobilization and impotence of their armies. The average allied strength in Africa in 1916 and 1917 was approximately 50,000 men. In that period there were over 100,000 hospital admissions for malaria from that force. During the Gallipoli campaign the greater part of the 120,000 medical casualties in the British Army were from dysentery. These illustrations supply ample evidence of the importance of tropical and parasitic diseases in warfare in the torrid zone and the necessity for adequate control measures.

The rapidity of movement of modern warfare greatly complicates this problem which is essentially protection of personnel against contaminated food and water, flies, and the attacks of numerous different arthropod vectors of disease including various mosquitoes, a variety of biting flies, the body louse, several species of ticks, and certain mites. Complete protection against many of the endemic diseases in the tropics is difficult if not impossible even under optimal peacetime conditions. The technique of mechanized warfare and the circumstances of actual combat are

such as largely to reduce protective measures to those which are directly applicable to the individual. Progress in the field of immunization, techniques of field sanitation, provision of protective uniforms, nets, and the utilization of efficient insecticides and repellants combine to render the hazards of tropical operations far smaller than might be anticipated.

Despite the utilization of all available measures, however, it is only reasonable to anticipate a higher sick rate than would be experienced in operations in the temperate zone, and that a variety of infections will be encountered, many of which are classed as parasitic diseases.

Bacterial infections do not properly belong in this category. The importance of bacillary dysentery as a war disease, however, is such that it cannot be passed over. The *Shigella* group is widespread over the world, especially where local conditions of sanitation are imperfect and permit direct contamination of food and drink by human excreta, or indirect contamination by flies having access to human feces. The Shiga bacillus particularly is widely endemic in the tropics and in the Orient. It produces far more serious disease and a much higher mortality rate than do the less toxic Flexner and Sonne-Duval strains.

Bacillary dysentery is important because of the serious acute disease, the not infrequent development of chronic dysentery following the acute phase, and the relatively high incidence of carrier states which may persist for considerable periods of time. *Shigella dysenteriae* produces an acute inflammation of the mucosa of the intestine, particularly the colon, with extensive ulceration, sloughing of the mucous membrane, hemorrhage, and severe secondary infection of the deeper tissues. Shiga bacillus infections, particularly, are commonly accompanied by profound toxemia, and in some epidemic outbreaks have been accompanied by mortality rates as high as 50 per cent. Clinically, acute bacillary dysentery is characterized by fever, toxemia, abdominal cramps, severe tenesmus, and stools which after the first few hours consist of little other than gelatinous masses of blood-stained mucus swarming with the bacilli. There are few diseases which present a greater hazard of infection to attendants and even indirect contacts.

It is impossible to evaluate the risk of chronicity, especially in view of the universal use of the sulfonamide drugs in the treatment of the acute stage. The British experience in the last war indicates that from 2 to 3 per cent became chronic. The clinical picture and the pathology of chronic dysentery are identical with those of chronic ulcerative colitis. Such individuals are seldom completely well and commonly suffer repeated acute recurrences over periods of many years. There is probably no disease which we have to meet in this war which carries so great a

potential threat of chronic invalidism and disability to such large numbers of individuals.

The magnitude of the carrier problem is likewise impossible to estimate. Again to cite British figures, in the last war approximately 3 per cent of recovered cases became persistent carriers. Actually this figure should probably be higher since the bacilli are recoverable from such individuals by culture only intermittently.

With the advent of the sulfonamide drugs, especially sulfathiazol, sulfaguanidine, and sulfadiazene, the means of treatment have been greatly strengthened. Some reservation is still necessary, however, since the efficacy of these drugs against toxic, virulent strains of the Shiga bacillus has not been finally evaluated. In the less severe infections produced by the Flexner and Sonne-Duval strains, the results of chemotherapy are not entirely clear-cut. Even in these milder cases chronic infections and carrier states are not eliminated.

Infection by *Endameba histolytica* presents an important problem not so much because a high incidence of acute dysentery is to be anticipated as because of the hazard of the late and dangerous complications. This protozoan parasite has a wide if not a universal distribution, despite the belief, long-held, that amebiasis is strictly a tropical disease. It is well established now that the limiting factors in the epidemiology of this infection are sanitation and hygiene—not climate. A further misconception lies in the older nomenclature. The term "amebic dysentery," still present in many medical texts emphasizes unduly a relatively uncommon clinical type. In fact, the infection is far more usual than such a descriptive term would seem to indicate. However, on the basis of past experience it would appear unlikely that much more than 10 per cent of the total cases of dysentery in this war will be attributable to *Endameba histolytica*.

This parasite produces ulceration of the colon, especially the cecum and proximal portion, which in many instances is of mild degree, and unaccompanied by symptoms of note. In others, the infection becomes progressive, the ulcers extend, becoming confluent and penetrating into the deeper layers, producing not only the clinical picture of acute dysentery, but at times exsanguinating hemorrhage, perforation, and death from peritonitis. In other instances the trophozoites penetrate into radicles of the portal vein, are carried by the blood stream to the liver where they initiate again progressive tissue necrosis and produce those very serious complications, amebic hepatitis or abscess of the liver. Liver abscess is said to occur in one of every eight cases of amebic dysentery. Prior to the formulation of modern treatment, this condi-

tion carried a mortality of 40 to 60 per cent, a rate which is still prevalent if the often difficult diagnosis is not made, or if improper therapy is used.

Amebiasis may persist unrecognized for many years producing a varied and varying symptomatology, often quite mild, but ultimately going on to serious hepatic involvement. The writer has seen an instance in which it appeared that the infection remained quiescent for 19 years. At the end of that period, however, a nearly fatal liver abscess occurred.

With emetine, the oxyquinoline drugs, and the organic arsenical carbarsone we have the armamentarium necessary for clinical cure, and equally important, eradication of the infection. This statement is valid, however, only if these drugs are used in full realization of the limitations of each and the basic pathology of the disease. Most of the problems and the dangerous situations result either from incorrect or too-long-delayed diagnosis, or inefficient utilization of the available drugs.

The amebae are situated on the surface of the mucous membrane of the colon and within the tissues as well. In the latter situation they are rapidly destroyed by emetine which is distributed by the blood stream. This drug does not destroy those on the surface. This fact accounts for Colonel Craig's observation that 85 per cent of cases treated by emetine alone relapse. The oxyquinoline group on the other hand, while lethal to amebae in the intestinal contents and on the surface of the mucosa, are not absorbed in sufficient concentration to exert an amebicidal action on those in the depths of the tissues. Carbarsone is intermediate in its pharmacological action. It follows logically that simultaneous combined treatment should give the best results and this is the fact in actual practice.

Frequently, however, diagnosis is difficult even of liver abscess. Not infrequently this complication develops "silently" and because of weight loss, low fever, secondary changes in the right lung, and cough, it is misdiagnosed as tuberculosis. In other instances it may masquerade convincingly as chronic malaria, or even kala azar.

By far the most important of the protozoal diseases, however, is malaria. The average practitioner in the temperate zone has no conception of the hazards of tropical malaria. The disease as it is seen in the cooler latitudes is produced by *Plasmodium vivax* which, although it may produce severe illness and marked anemia and disability, is not attended by the grave prognosis and often fatal complications of the tropical form. In part this difference is probably attributable to the fact that *P. vivax* attacks the immature red cells or reticulocytes and not the mature erythrocytes. *Plasmodium falciparum*, the cause of the tropical malaria, on the other hand attacks all forms of the red blood

cells and unless checked rapidly produces an overwhelming parasitemia never encountered in the other forms of malaria. Furthermore *P. falciparum* seems to produce changes in the physical characteristics of the infected cells. The cell membranes become "sticky" and the parasitized cells tend to adhere to capillary endothelium, and to each other producing both emboli and thrombi particularly in the brain and the viscera, and leading to ischemia and anoxemia.

Falciparum malaria is the predominant form throughout the tropics and in many areas in the subtropics. The rapidity of progression and the variability and severity of the clinical picture are to be attributed in part to the rapid blood destruction incidental to the high parasitemia, and in large part to the plugging of visceral capillaries. In the face of such a mechanism it is not surprising that the clinical picture is often confusing. The frequency with which certain structures are affected has given rise to a rough classification of the clinical types of disease. Cerebral localization, which is common, may give rise to the rapidly fatal hyperpyrexial form with death occurring within a very few hours, the delirious and comatose type, and other symptom complexes indicative of profound disturbance of the central nervous system. Predominant localization in the gastrointestinal tract may produce the bilious-remittent or gastric types characterized by profuse and continuous vomiting, or the frequently fatal afebrile algid type which is accompanied by diarrhoea so severe as to resemble cholera. In other instances the symptom complex may present many of the features of acute bacillary dysentery with abdominal pain, frequent stools containing mucus and blood in which numerous parasitized erythrocytes may be demonstrated. Not infrequently *falciparum* malaria may closely mimic acute appendicitis.

The hazard of this infection is still further augmented by its association with hemoglobinuric fever or blackwater fever. Although the exact etiology is unknown, intermittent quinine therapy, exposure, chilling, and exhaustion are generally regarded as precipitating causes. The writer has seen a fatal instance occur within two hours of the administration of oil of chenopodium for the elimination of a moderate load of *Necator*.

Blackwater fever usually occurs in individuals who have a latent malaria of some standing. Characteristically the onset is acute with chill, sharp rise in temperature, and profuse vomiting accompanying sudden massive intravascular hemolysis. This phenomenon is followed shortly by the passage of dark-colored or even black urine due to the presence of hemoglobin and its derivatives, and deepening jaundice. The acute blood destruction produces a rapidly developing and grave

anemia. In severe cases the red cell count may fall to two million per cubic millimeter or even lower within 24 hours. The urine is acid, and in addition to the blood pigments, contains large amounts of albumin and casts. In the progressive case, especially if the urine is not rendered alkaline, mechanical plugging of the renal tubules occurs leading to anuria, nitrogen retention, and death from renal failure. In other fatal instances, hemolysis ceases, the urine clears, and convalescence may seem established for some days only to be terminated by abrupt recurrence of hemolysis eventuating in death.

Augmenting the hazard of *falciparum* malaria, especially in the face of its complications, is the not uncommon difficulty in arriving at an accurate diagnosis. The clinical picture frequently bears no remote resemblance to the classic malaria of the text books. The so-called characteristic tertian fever is seldom present. Unlike *Plasmodium vivax* infection, in severe cases the *Plasmodium falciparum* may be demonstrable only in thick blood films and may be missed entirely if only thin films are used. This pitfall constitutes a paradox seldom if ever appreciated by physicians who have not had special training and experience. It is probably related to the sporulation in the capillaries of the viscera rather than in the peripheral blood. Some degree of involvement of the central nervous system is frequent, and the appearance of drunkenness is a very common mode of onset of cerebral malaria. The practice of Sir Patrick Manson, carried on at the Seamens Hospital in London, is worth serious consideration by all who may come into contact with clinical malaria—in the presence of intoxication treat for malaria first and then for alcoholism.

Quinine has for long been regarded as the only truly effective agent in the therapy of malaria. This widespread conviction has given rise to much unwarranted concern about the effects of the shortage resulting from the war. Fortunately we have a satisfactory substitute in atabrine. In fact some authorities regard it as a more efficient antimalarial drug. It is a yellow dye which is excreted slowly by the kidneys. It stains the tissues yellow producing a spurious "jaundice." Like quinine it may be administered by mouth or parenterally. When taken by mouth it should be given immediately after meals or accompanied by a sweetened drink as some individuals have symptoms of irritation of the gastro-intestinal tract when it is taken on an empty stomach. This drug has had very extensive trial over a number of years and it is well established that it is an effective prophylactic and suppressive agent as well. In some respects it is preferable to quinine for this purpose for military personnel. A third standard antimalarial drug is plasmoquine. How-

ever, it is effective only against the gametocytes or sexual forms of the parasites and consequently is not of value in the treatment of the acute disease. Administration following quinine and atabrine therapy seems to lower the incidence of late recurrences.

Of the other protozoan parasites of man the *Leishmania* are the most important. Visceral leishmaniasis is endemic in many of the areas of military operations. Kala azar occurs in the Mediterranean basin, the Sudan, the Near East, India, and China. The cutaneous leishmaniases have a very similar distribution.

It is impossible to estimate how much of a problem these conditions may present as the epidemiology is not entirely clear. They are co-extensive with certain sandflies of the genus *Phlebotomus* and there is evidence to indicate that at least three members of the group, *P. argentipes*, *P. papatasii*, and *P. chinensis* may transmit the infection under experimental conditions.

The *Leishmania* localize in the cells of the reticulo-endothelial system, multiplying in them and greatly distending and distorting them. The visceral forms of the disease are characterized by prolonged irregular fever, chronicity, splenic and often hepatic enlargement, emaciation, anemia and leukopenia. In the cutaneous form the infection is usually localized especially to exposed skin areas where nodules and ulceration result. These lesions are seldom accompanied by involvement of the viscera. Dermal leishmaniasis may also occur as a complication or sequel of kala azar especially in inadequately treated cases. Various preparations of antimony, notably certain pentavalent compounds, have been used successfully in the therapy of these conditions.

It is unlikely that metazoan parasites will present to the armed forces problems approaching those of the Protozoa. Diseases produced by these agents tend to be less acute, slower of development, more chronic, and in many instances a considerable period of time is required between infection and the development of a severe pathologic response. There are of course important exceptions to this. The geographical distribution of certain of the important helminths is quite restricted in some instances. In others the particular epidemiology is such as to present no great obstacle to the institution of adequate protective measures. Certain ones, however, have wide geographical distributions and utilize insect vectors which make protection difficult and at times impossible.

Although hookworm infestation will undoubtedly occur, it is improbable that it will be accompanied by hookworm disease. The clothing and shoes worn by troops and the strictly local soil contamination will minimize both the number and the intensity of exposures. It is a somewhat

different matter with *Ascaris lumbricoides* since it does not pass through a free-living phase in its life cycle and since its epidemiology is essentially that of amebic and bacillary dysentery—transmission by fecal-contaminated food and water, and by flies. One would not anticipate, however, a significant morbidity rate.

With the Filaroidea the situation is quite otherwise. In certain areas, unless there is rigid water discipline, infestation by the Guinea worm, *Dracunculus medinensis*, will occur. Some months following ingestion of water containing infected *Cyclops* the adult female worm reaches the skin surface usually on the lower extremities producing a local inflammatory reaction and at times an accompanying toxemia. Finally a blister forms, ruptures, and reveals the opening of the burrow in which the anterior portion of the worm lies and through which living larvae are discharged in response to the stimulus of contact with water. Secondary bacterial infection is a serious hazard, but even in the absence of infection considerable disability is produced.

Another member of this group, *Onchocerca volvulus*, is widely distributed through equatorial Africa and is transmitted by at least five species of flies belonging to the genus *Simulium*. It produces subcutaneous fibrous tumors which contain the adult worms, and at times visual disturbances progressing to blindness. The geographical distribution, which is outside the present and probable future zones of combat, makes it unlikely that this parasite will prove to be of importance in war medicine.

Wuchereria bancrofti, however, will almost certainly be of some importance in the Pacific campaign. It is widely prevalent throughout the entire tropical zone and many of the islands of the southwest Pacific are heavily infested. A large number of mosquitoes including certain *Aedes*—among them *Aedes aegypti*, *Anopheles*, *Culex*, and *Mansonia*—transmit the infection in these areas. This parasite, by causing a progressive sclerosing and fibrosing lymphadenitis and lymphangitis of the regional lymphatics which are in anatomical relationship to the site of the adult worms, produces endemic or tropical elephantiasis of the legs, the scrotum, less often the upper extremities and the breast, orchitis, hydrocele, and chyluria or milky urine by obstruction of the thoracic duct and secondary rupture of the lymphatic vessels in the wall of the bladder. Filarial fever and acute lymphangitis are not uncommon. Unfortunately there is no specific therapy.

The trematodes *Schistosoma mansoni*, *Schistosoma hematobium*, and *Schistosoma japonicum* are important disease-producing agents and are distributed in certain of the theaters of operations. *S. mansoni* is en-

demic in the Nile valley, much of central Africa, parts of South America especially the northern coast, and certain of the West Indies islands including Puerto Rico. Transmission depends upon exposure of the skin to water containing the infective cerceriae and consequently effective water discipline including avoidance of wading and bathing in untreated water confers complete protection. However, this may not be practicable at all times. Six to eight weeks after infection the worms have reached maturity in the radicles of the portal system and oviposition takes place in the finer venules leading to ulceration into the lumen of the colon, producing dysentery. Ova likewise accumulate in the liver in heavy infections producing a periportal type of cirrhosis with secondary splenomegaly and ascites.

S. hematobium has a wide distribution in Africa and also occurs in areas in the Near East and in western Asia. The adults of this species invade the pelvic veins, especially the vesiculo-prostatic, the pubic, and the uterine plexuses. Ova are deposited in the mucosa of the bladder where they produce ulceration, cystitis, papillomata, and a variety of disturbances of the genito-urinary system, and at times involvement of the rectum as well.

S. japonicum is confined to the Far East. Infection by this parasite is characterized by chronic dysentery, great enlargement of the liver and spleen, and finally cirrhosis of the liver and ascites. The intravenous administration of tartar emetic has proved of great value in the treatment of all forms of schistosomiasis. The trivalent antimony compound Fouadin is less toxic and perhaps even more effective.

Even such a cursory consideration of the tropical and parasitic conditions to which military and naval personnel will be exposed demonstrates the magnitude of the problems confronting the medical services. The experience of the Allied Armies in Macedonia, Gallipoli, and Africa in the last war shows conclusively how vitally and sometimes tragically these conditions may affect the outcome of a campaign. It is likewise apparent that the distribution of these infections, and factors such as the season of the year which may affect natural transmission rates must, whenever possible, weigh heavily in the decisions of staffs planning campaigns and the major strategy of the war. It is likewise evident how great is the need in troop units and hospital installations for medical officers trained in tropical medicine and medical parasitology. Fortunately this problem was visualized by the medical departments of the Army and Navy long before our entry into the war. At the request of the Surgeons General the National Research Council set up, among others, a subcommittee on Tropical Diseases which has been acting

continuously in a consultative capacity. In the summer of 1940 a special course in Tropical Medicine was established at the Army Medical School with an instruction staff including many distinguished scientists and physicians At the present time it is graduating some two hundred medical officers every eight weeks. In this way the Army and the Navy are compensating for the failure of our educational institutions to recognize earlier the importance of tropical and parasitic medicine.

We can face the medical hazards of the war with no inconsiderable assurance. Although morbidity rates may prove to be higher than in a conflict limited to the temperate zone, we can be assured that we shall not be immobilized and rendered impotent as occurred to armies in the last war.

The implications of tropical and parasitic medicine, however, extend beyond the duration of the war and intimately concern the population of the United States. It is inevitable that numerous carriers and individuals with latent infections will be scattered over the country following demobilization. Certain of these conditions will be transmissible and will subsequently appear in persons who have not been out of the country. Furthermore, the clinical picture attending the combination of familiar endemic disease such as pneumonia with one of these less familiar parasitic infections may be most bizarre and atypical. Each obscures the characteristic features of the other. Contrary to the classic dictum of medical teaching against multiplicity of diagnoses, it must be recognized *a priori* that such multiplicity will occur and must be promptly recognized. There is urgent need for the inclusion of much more parasitology and tropical medicine in the curricula of our medical schools.

DISCUSSION OF THE PAPER

Dr. W. Oliver (*Long Island College of Medicine, Brooklyn, N. Y*):

Parker recently showed that ticks could be used as a medium for bringing in fevers for which the species is not normally a host. If experimentally infected, these ticks remain infected for 33–46 days.

Dr. G. C. Shattuck (*Harvard Medical School, Cambridge, Mass.*):

Resistance is a factor of great importance in these diseases. You can not cure malaria by drugs alone. Drugs plus the bodily developed resistance is needed. Anything that will improve the general condition of the patient is a help in recovery.

In treatment of elephantiasis caused by filaria, some good was obtained by vaccination. There was no effect on the parasite, but the treatment strengthened the resistance of the patient.

We do not know the limits of distribution but we do know that the limits are spreading.

Dr. McIver Woody (*Standard Oil Company of New Jersey, New York, N. Y.*):

In tropical commercial projects research has been laid aside temporarily. You can sell to the oil business medical findings which save man hours by preventing diseases.

Dr. H. E. Meleney (*New York University Medical School, New York, N. Y.*):

In helping to outline a new course in parasitology for medical schools I believe that principles of biology could well be taught from a study of parasites and their vectors rather than the more conventional entomology.

Planes may bring in persons with incipient autumnal malaria which is still in the incubation period and therefore undetected. The affected person may go into coma suddenly. Many New York doctors will fail to diagnose this disease, thus depriving the patient of prompt treatment – so essential for success.

Dr. R. Matheson (*Cornell University, Ithaca, N. Y.*):

Medical schools should insist on entomology and parasitology as part of the entrance requirement. If they did so, colleges would be glad to introduce these subjects.

Annals of The New York Academy of Sciences
Volume XLIV, Art. 4. Pages 263-443
November 12, 1943

HIGH POLYMERS*

By

Raymond M. Fuoss, J. Abere, W. O. Baker, Henry Eyring, John D. Ferry, Paul J. Flory, C. S. Fuller, G. Goldfinger, R. A. Harman, Maurice L. Huggins, H. M. Hulburt, H. Mark, H. Naidus, Charles C. Price, John Rehner, Jr., Robert Simha, and A. V. Tobolsky

CONTENTS

*This series of papers is the result of a conference on High Polymers held by the Section of Physics and Chemistry of The New York Academy of Sciences, January 8 and 9, 1943.

Publication made possible through grants from the Conference Publication Revolving Fund and the income of the Permanent Fund.

INTRODUCTION TO THE CONFERENCE ON HIGH POLYMERS

By Raymond M. Fuoss

From Research Laboratory of General Electric Company, Schenectady, New York

Although the field of high polymers is a comparatively new subject, the technical importance of these compounds has led to a very rapid development of empirical knowledge concerning them. Fundamental work, from the academic point of view, has also made considerable progress. Roughly speaking, the polymer chemist in the technical field is contented when he has a resin or plastic which performs satisfactorily under a given set of conditions and which can be made at a reasonable price. The academic worker, on the other hand, is not particularly interested in the industrial applications of the compounds he is studying, but for his part, he is not happy until he knows (or at least, thinks he knows) why they behave as they do, and why a given collection of atoms has a given reproducible set of properties. This coexistence of two kinds of knowledge, acquired as the result of entirely different mental urges, has turned out to be very valuable in the field of polymer chemistry, in that there has been a mutual stimulation of applied and fundamental research. Problems have arisen in practical work which have led to purely academic studies. In compensation, suggestions based on seemingly abstract topics have found application in technology.

This conference was planned in order to present a review of the current status of our knowledge of polymers of high molecular weight, and to report recent work on a number of important problems within the field. A further purpose of the conference was to provide an opportunity for open discussion among a group, all of whom are actively working on one or another phase of polymers. The final (expanded) drafts of the papers will be submitted after the meeting, so that contributions made during discussion can also be included in publication. For this reason, some of the preprints are tentative or merely introductory in form and intention.

The first question to be considered logically concerns the formation of macromolecules. The mechanism of reaction, in the stages of chain initiation, propagation and termination; the role of catalysts and the mechanism of their part in starting polymerization; the correlation between monomer structures and their susceptibility to polymerization catalysts; the structure of the final product, and the distribution of

molecular weights; the kinetics of the various stages of the polymerization reactions; all these detailed problems are involved. A number of points of attack are possible; a combination of the methods of the physical and organic chemist, together with help in interpretation from the theoretical side, gives the most effective results. The present program was organized with the intention of combining these different points of view so that, for example, reaction mechanism will be discussed from both the experimental and theoretical aspects, evidence concerning structure as obtained by the methods of organic chemistry and by X-ray technique will be presented, and so on, in the hope that a final clearer picture will emerge than could have been obtained from one perspective alone.

Our opening paper deals with the general theory of polyreactions. Other papers on the program will treat special problems in the fields of kinetics and structure, and correlations between structure and properties. Since, in last analysis, the technical importance of high polymers is largely due to their unique mechanical properties, two papers presenting experimental and theoretical treatments of flow properties are included. As in the classical field of low molecular weight compounds, the colligative properties of high polymers give information concerning their molecular weight and also on forces between molecules. As might be expected, the classical theory requires special adaptation to the field of high polymers in order that a reliable interpretation of experimental results may be made. The limiting law for osmotic pressure and the deviations as functions of concentration are the subject of our final paper.

RECENT RESULTS ON THE KINETICS AND ELEMENTARY STEPS OF POLYREACTIONS

By

J. Abere, G. Goldfinger, H. Mark, and H. Naidus

From Brooklyn Polytechnic Institute and Queens College, New York, N. Y.

INTRODUCTION

If one wants to follow quantitatively the course of a polymerization or polycondensation reaction, one has first of all to carry out certain measurements. Usually the following quantities are recorded.

(a) *The total amount of monomer which has been converted into polymer (dimer and higher) as a function of time, nature of catalyst and solvent, monomer concentration and temperature.* This corresponds to the weight of polymer (of all degrees of polymerization) formed at any instant. Particularly important is the *initial rate* with which the polymer is produced (the monomer disappears), because, in the early stages of the reaction, the conditions are usually not yet too complicated and can, with some success, be expressed by appropriate rate equations. There are many experimental methods available to determine the rate of polymer formation in each individual case,[1] [2] [3] but it would lead us too far afield to enumerate them in this article.

(b) *The number or weight average degree of polymerization of the polymer produced as a function of time and of the other experimental parameters mentioned above, such as amount and nature of solvent, concentration and character of catalyst or inhibitor, temperature, etc.* It is well known that a reliable determination of these quantities at present still meets with considerable uncertainties and difficulties,[4–8] but it seems that osmotic pressure measurements in very dilute solution provide a comparatively safe way to secure the number average for the degree of polymerization, if the membrane is such that it does not let too much of the lower molecular weight

[1] Burk, R. E., Thompson, H. E., Weith, A. J., & Williams, I. 'Polymerisation Reinhold New York. 1937.
[2] Carothers, W. H. "Collected Papers." Interscience Publishers. New York. 1940.
[3] Mark, H., & Raff, R. "Highpolymeric Reactions." Interscience Publishers. New York 1941
[4] Bartovics, A. Ph. D. Thesis, Polytechnic Institute of Brooklyn. 1942.
[5] Fordyce, R., Lovell, E. L., & Hibbert, H. Jour. Am Chem Soc. 61: 1905. 1939
[6] Fuoss, R. M., & Mead, D. J. Jour. Phys Chem 47. 59. 1943
[7] Huggins, M. L. Jour. Phys. Chem. 46: 151. 1942, Ann. N. Y. Acad. Sci 44: 431. 1943
[8] Kemp, E. A., & Peters, H. Ind. Eng Chem. 34: 1097. 1942.

material pass through. The weight average can be determined in principle[3, 4] by diffusion and ultracentrifuge, the viscosity average by viscosity measurements. But the first method often offers experimental difficulties, while the others allow at best only a rough determination of the order of magnitude of the degree of polymerization.

In simple cases, measurements of the type (a) and (b) are sometimes sufficient to work out a rather detailed picture of how the polymer molecules are formed by interaction of the monomer with itself and with intermediate stages of the final product. Under more complicated conditions, however, they are not sufficient. In such cases it is very advantageous to carry out additional studies in order to obtain the necessary information for a proper description of the reaction mechanism. Such additional experiments are the following.

(c) *Determination of the molecular size distribution curve for the polymer at different times and under different experimental conditions, such as temperature, the use of a certain type of catalyst, etc.* No thorough investigation of this kind has yet been published, although some of P. J. Flory's,[9–13] R. Simha's[14] and G. V. Schulz's papers[15–18] contain very interesting steps in this direction. More recently T. Alfrey[19] and A. Bartovics[4] have started systematic distribution curve measurements on samples of polystyrene which had been prepared under different conditions, and J. Abere, G. Goldfinger and H. Naidus[20] correlated these studies with kinetic measurements of catalyzed styrene polymerization under rather widely varied conditions.

(d) *Chemical analysis of the polymer at different stages of its formation, mostly to find out whether fragments of an inhibitor, a catalyst, or the solvent have been incorporated in the long-chain molecule of the polymer.* Carothers[2, 21] and Staudinger[22–24] have been interested

---

[9] **Flory, P. J.** Jour. Am. Chem. Soc. **58**: 1877. 1936.
[10] **Flory, P. J.** Jour. Am. Chem. Soc. **59**: 241. 1937.
[11] **Flory, P. J.** Jour. Am. Chem. Soc. **61**: 1518, 3334. 1939.
[12] **Flory, P. J.** Jour. Am. Chem. Soc. **62**: 1057, 1561, 2225, 2261. 1940.
[13] **Flory, P. J.** Jour. Am. Chem. Soc. **63**: 3083, 3091, 3096. 1941.
[14] **Ginell, R., & Simha, R.** Paper presented at the Buffalo meeting of the American Chemical Society in September, 1942. Jour. Am. Chem. Soc. **65**: 706, 715. 1943.
[15] **Schulz, G. V.** Zeit. physikal. Chem. **B30**: 399. 1935; **B32**: 27. 1936.
[16] **Schulz, G. V., & Husemann, E.** Zeit. physikal. Chem. **B34**: 187. 1936; **B36**: 184. 1937, **B39**: 246. 1938.
[17] **Schulz, G. V., Dinglinger, A., & Husemann, E.** Zeit. physikal. Chem. **B43**: 25, 47, 385. 1939; **B50**: 306. 1941.
[18] **Schulz, G. V., & Wittig, G.** Naturwiss. **27**: 387, 456, 659. 1939.
[19] **Alfrey, T.** Ph. D. Thesis, Polytechnic Institute of Brooklyn. 1942.
[20] **Abere, J., Goldfinger, G., Naidus, H., & Mark, H.** Paper presented at the Buffalo meeting of the American Chemical Society in September, 1942.
[21] **Carothers, W. H.** Chem. Rev. **8**: 402. 1931; Trans. Faraday Soc. **32**: 44. 1936.
[22] **Staudinger, H., & Schulz, G. V.** Ber. **68**: 2320. 1935.
[23] **Staudinger, H., & Steinhofer, H.** Ann. **517**: 35. 1935.
[24] **Staudinger, H.** Trans. Faraday Soc. **32**: 97. 1936.

in such kind of analysis, although they did not utilize their findings for kinetic considerations. More recently J. Abere,[20] S. Abkin,[25, 26] H. N. Alyea,[27] J. L. Bolland,[28] J. W. Breitenbach,[29–31] R. E. Burk,[32, 33] A. Dinglinger,[17] D. A. Durham,[34] J. J. Gartland,[27] G. Goldfinger,[20] H. R. Graham,[27] E. Husemann,[16] Lankelma,[33] S. Medvedev,[25, 35, 36] H. Naidus,[20] H. Pfann,[37] C. C. Price,[34, 39–41] G. V. Schulz,[16, 17] and G. Whitby [42–44] have analyzed polymers of comparatively low molecular weight and succeeded in locating, on one or both ends, parts of catalyst or solvent molecules and have drawn fairly far-reaching conclusions as to the course of the reaction. Quite recently such experiments have been extended into the domain of higher degrees of polymerization by the use of radioactive isotopes (bromine) by H. Pfann [37] and G. D. Salley.[38]

Physical methods (Raman and infra-red spectroscopy) have also been applied to the investigation of the end groups and to determine the progress of polymerization with time. E. Briner,[45] I. Inoue,[45a] S. Mizushima,[45a] D. Monnier,[45] I. Morino [45a] and B. Sucz [45] have succeeded in following the disappearance of the aliphatic double bond during styrene polymerization using the intensity changes of certain Raman lines. Finally, C. S. Marvel and his collaborators [47–52] have used optical activity with great success in following the course of vinyl-type co polymerizations.

(c) Very important information as to the different steps of a polymerization reaction can be obtained if one succeeds in *counting the number of active centers which start the propagation of the chains.*

---

25 **Abkin, S., & Medvedev, S.** Trans. Faraday Soc. **32**: 286. 1936
26 **Mamontora, O., Abkin, S., & Medvedev, S.** Acta physicochim. USSR **12**: 269. 1940
27 **Alyea, H. N., Gartland, J. J., & Graham, H. R.** Ind. Eng Chem **34**: 458. 1942.
28 **Bolland, J. L.** Proc. Roy. Soc. **A178**: 24. 1941.
29 **Breitenbach, J. W., & Raff, R.** Monatsh. **69**: 1107. 1936
30 **Breitenbach, J. W., & Rudorfer, H.** Monatsh. **70**: 37. 1937.
31 **Breitenbach, J. W.** Monatsh. **71**: 275. 1938. Zeit physikal. Chem. **B45**: 101. 1939.
32 **Burk, R. E., Baldwin, B. G., & Whitacre, C. H.** Ind. Eng. Chem. **29**. 326. 1937.
33 **Burk, R. E., Laskowski, L., & Lankelma, H P.** Jour. Am Chem. Soc. **63**: 3248. 1941.
34 **Price, C. C., & Durham, D. A.** Jour. Am. Chem. Soc. **64**: 2508 1942
35 **Chilikina, E., & Medvedev, S.** Acta physicochim. USSR **12**. 203. 1940
36 **Kamenskaja, S., & Medvedev, S.** Acta physicochim. USSR **13**: 565. 1940
37 **Pfann, H.** Master's Thesis, Polytechnic Institute of Brooklyn. 1942.
38 Private communication. Comp. also **Kern, W., & Kammerer, H.** Jour prakt Chem. [2] **161**: 100. 1942.
39 **Price, C. C., & Kell, R. W.** Jour. Am. Chem. Soc. **63**: 2798. 1941.
40 **Price, C. C., Kell, R. W., & Krebs, E.** Jour. Am Chem. Soc. **64**: 1103. 1942.
41 **Price, C. C.** Ann. N. Y. Acad Sci. **44**: 351. 1943.
42 **Whitby, G. S., & Katz, M.** Jour. Am. Chem. Soc **50**: 1160. 1928.
43 **Whitby, G. S., & Crozier, R. M.** Can. Jour. Res. **6**: 203. 1932.
44 **Whitby, G. S.** Trans. Faraday Soc. **32**: 315. 1936.
45 **Monnier, D., Suez, B., & Briner, E.** Helv. **21**: 2340. 1936.
45a **Mizushima, S., Morino, I., & Inoue, I.** Chem. Soc. Japan **12**. 136. 1937.
46 **Moore, J. K., Burk, R. E., & Lankelma, H. P.** Jour. Am. Chem. Soc. **63**: 2954. 1941.
47 **Marvel, C. S., & Levesque, C. L.** Jour. Am. Chem. Soc. **60**: 280. 1938.
48 **Marvel, C. S., & Denoon, C. E., Jr.** Jour. Am. Chem. Soc. **60**: 1045. 1938.
49 **Marvel, C. S., Sample, J. H., & Roy, M. F.** Jour. Am. Chem Soc. **61**: 3241. 1939.
50 **Marvel, C. S., & Cowan, J. C.** Jour. Am. Chem Soc. **61**: 3156. 1939.
51 **Marvel, C. S., Dee, J., & Cooke, H. G.** Jour. Am. Chem. Soc. **62**: 3499. 1940.
52 **Marvel, C. S., Jones, G. D., Mastin, W. T., & Schertz, G. L.** Jour. Am Chem. Soc. **64**: 2356. 1942.

This has been done by H. W. Melville and collaborators,[53, 54] J. L. Bolland,[55] T. T. Jones[56] and R. F. Tuckett[57] in a very interesting series of recent articles. Similar insight in the mechanism can be obtained if it is possible to *localize* geometrically the initiation reaction and to observe the gradual growth of the polymer from these starting points or starting areas. J. Abere,[20] S. Abkin,[25, 26] J. L. Bolland,[28] C. B. Davies,[58] G. Gee,[58, 59] G. Goldfinger,[20] W. Jorde,[60] S. Medvedev,[25, 61] H. W. Melville[54] and H. Naidus[20] have performed experiments of this kind which have greatly helped to advance our knowledge of the different elementary processes involved. Particularly, the studies of localized polymerization in the gaseous phase promise to be of value for a better understanding of the chemical nature of these elementary steps.

These are the main methods which have been, and are being, applied to provide a sound experimental basis for theoretical considerations and speculations as to the chemical nature of the individual steps of polycondensation and polymerization reactions, and as to how these elementary processes cooperate in building up the large molecules of the polymer.

The first step in the evaluation of the experimental data is mostly concerned with determining the nature of what one usually calls the *formal kinetics* of the composite reaction in question. We look for a set of equations to describe how the elementary reaction steps take place simultaneously or consecutively, how they consume the monomer, how they produce certain intermediate (frequently short lived) configurations and finally build up the stable end product with all its characteristic properties.

Different attitudes can be taken in attempting to work out this formal kinetics of polyreactions. One method, which has been applied with great success by various authors,[33, 36, 46, 53, 62–64] consists in making certain probable assumptions as to the elementary processes involved, and then writing down differential rate equations which represent these col-

[53] Melville, H. W. Trans. Faraday Soc. 32: 258. 1936; Proc. Roy. Soc. A166: 163, 511. 1937; A167: 99. 1938

[54] Melville H. W. Trans. Inst. Rubber Ind. 15: 209. 1939.

[55] Melville, H. W., & Bolland, J. L. Oest. Chem. Ztg. 42: 201. 1939.

[56] Melville, H. W., & Jones, T. T. Proc. Roy. Soc. A175: 392. 1940.

[57] Melville, H. W., Jones, T. T., & Tuckett, R. F. Chem. Ind. 59: 267. 1940.

[58] Gee, G., Davies, C. B., & Melville, H. W. Trans. Faraday Soc. 35: 1298. 1939

[59] Gee, G. Trans. Faraday Soc. 34: 712. 1938; 36: 1171. 1940.

[60] Unpublished experiments.

[61] Medvedev, S., Chilinkina, E., & Klimenkov, V. Acta physicochim. USSR 11: 751. 1939.

[62] Norrish R. G. W., & Brookman, E. F. Proc. Roy. Soc. A163: 205. 1937; A171: 147. 1939 Trans. Faraday Soc. 35: 1087. 1939.

[63] Suess, H., Pilch, K., & Rudorfer, H. Zeit. physikal. Chem. A179: 361. 1937.

[64] Suess, H., & Springer, A. Zeit. physikal. Chem. A181: 81. 1937.

laborating steps. Integration of those equations leads then to expressions which can be compared directly with the experiments. If this check is satisfactory, it shows that the underlying assumptions are capable of giving a correct representation of the special polyreaction under consideration. This method has the advantage of being simple and flexible and of usually involving only moderately complicated mathematical considerations; it has the drawback that one has to make separate assumptions in each single case, which involves a certain risk, inasmuch as the final result depends on the particular choice of these assumptions.

The other method aims to develop once and for all sets of *general* differential equations for the main types of polyreactions, such as polycondensations, polymerizations, polydegradations or polyexchange reactions. These sets are integrated and every individual reaction can then be treated as a special case of one of the general types.[9–14, 31, 65–70] As long as only a few polyreactions were known, it did not seem worth while to develop a comparatively elaborate theoretical framework, but, at present, when one knows a large number of such reactions and wants to have a reliable platform to compare them in all details, it seems that this more systematic and general approach should be recommended. It has been pointed out that the algebraic expressions which one obtains as a result of the integration of the general sets of differential equations are somewhat clumsy and unwieldy It is true that they appear so in their general form, but as soon as one applies them to a special case they become rather simple and actually coincide with the expressions which have been developed separately in each individual case.

The development of a formal reaction kinetics, however, is only a preliminary and necessary step in the theoretical evaluation of experimental results, but not the final goal, which is, rather, a thorough understanding of the *chemical nature of each single elementary step involved.* To arrive at this, it seems advantageous to use all the results of the formal kinetics and to assign to each individual reaction (initiation, propagation, termination, chain transfer, etc.) a certain order of reaction, a definite frequency constant and a characteristic activation energy. If one succeeds in this way to describe *quantitatively* the various elementary steps, one has a fair chance to narrow down very materially the speculations as to their chemical nature.

---

[65] Chalmers, W. H. Can. Jour. Res. 7: 113. 1932, Jour. Am. Chem. Soc. 56: 912. 1934.
[66] Dostal, H., & Mark, H. Zeit. physikal. Chem. B29: 299. 1935, Trans. Faraday Soc. 32: 54 1936; Ind. Eng. Chem. 29: 595. 1937.
[67] Dostal, H., & Raff, R. Zeit. physikal. Chem. B32: 117. 1936.
[68] Iranyi, P. Jour. Am. Chem. Soc. 60: 2106. 1938; 61: 1784. 1939, 62: 2690. 1940.
[69] Mark, H., & Raff, R. Zeit. physikal. Chem. B34: 275. 1936.
[70] Wall, F. T. Jour. Am Chem. Soc. 62: 803. 1940; 63: 821, 18

It will be seen in the following that in some comparatively simple cases this has been achieved with good success, while in others we are still far from a satisfactory understanding of all influences which contribute to the formation of the final product.

After this general introduction, it seems appropriate to proceed now to the description of the various types of polyreactions and it appears logical to start with the simplest case, namely, with reactions in which only one single type of elementary step is involved.

POLYREACTIONS CONSISTING OF A SINGLE ELEMENTARY STEP

This elementary step repeats itself during the course of the reaction again and again under practically identical conditions and with very much the same rate. It acts between monomers as well as between chains which have already a certain length, without being sensibly affected by the chain length of the molecule to which the interacting groups belong. This type of reaction can be represented by

$$m_i + m_j \rightarrow m_{i+j},$$

where both i and j can assume any value from unity to very large values. W. H. Chalmers,[65] H. Dostal,[66] P. J. Flory [9] and R. Raff [3] have developed equations which take care of such multistep reactions, and Flory [12, 14] has investigated particularly whether the assumption of a homogeneous process having one single characteristic rate constant is in agreement with experiment. He found that polyesterifications follow this scheme and it is probable that most linear polycondensation products, such as polyesters, polyethers, polyamides, etc., are formed by processes of this type. In some cases, however, R. H. Kienle, P. A. van der Meulen and E. F. Petke [71] apparently have encountered more difficult conditions.

To describe the system under consideration before the reaction starts, we consider the polyesterification of an ω-hydroxycarboxylic acid and introduce the following symbols:

Total number of mols of the monomer	N_0
Weight of one mol of the monomer in grams	M_0
Total weight of the system in grams	$M = N_0 \cdot M_0$

After the reaction has proceeded for a certain time, t, a certain number of the OH groups and an equal number of the COOH groups will have reacted and a certain number of chains of various lengths will have been formed. To characterize our system we need now the following symbols:

[71] Kienle, R. H., van der Meulen, P. A., & Petke, E. F. Jour. Am. Chem. Soc. 61: 2258, 2268. 1939; 62: 1053. 1940.

Number of reactive or functional groups which have not yet reacted and are therefore still present	$N = N_0(1 - p)$
Number of functional groups which have reacted and therefore have disappeared	$N_0 - N = N_0 p$
Extent of reaction	$p = \dfrac{N_0 - N}{N_0}$
Number of members of an individual chain molecule	x
Number of x-mers in the system	N_x
Number fraction of x-mers	$m_x = \dfrac{N_x}{N_0}$
Weight of material comprised in x-mers	$s_x = N_x \, x \, M_0'$ †
Weight fraction of material comprised in x-mers	$M_x = m_x \, x$
Reaction constant for all steps	k

With this notation the differential equations of the polycondensation reaction assume the following form:

$$\frac{dm_1}{dt} = -km_1 \sum_1^{\infty} m_s. \tag{1}$$

This equation describes the disappearance of the monomer m_1, due to its reaction with itself and with chains of all possible lengths. The summation runs from $s = 1$ to $s = \infty$, and according to the above assumptions, the same rate constant, k, holds for all terms of this sum.

$$\frac{dm_x}{dt} = \tfrac{1}{2}k \sum_1^{x-1} m_s \, m_{x-s} - km_x \sum_1^{\infty} m_s. \tag{2}$$

This relation accounts for the formation and disappearance of the x-mer, m_x. The first term states that x-mers are formed by the combination of two smaller molecules, whose indices just add up to x; e.g., by one tetramer plus one (x-4)mer. If one carries out this summation, however, one counts each molecule, with s smaller than ($x - 1$), twice and hence has to divide this term by two. The second term describes the disappearance of the x-mers due to their reaction with any molecule from $s = 1$ to $s = \infty$, and is completely analogous to expression (1).

The solution of equations (1) and (2) is

$$m_x = p^{x-1}(1 - p), \tag{3}$$

† M_0' is the weight M_0 of the monomer minus 18 (weight of one mol water).

$$p = \frac{kt}{2 + kt} \tag{3a}$$

The total amount of the condensate from the dimer up to the most highly condensed molecules is given by

$$\sum s_x = Mkt\frac{4 + kt}{(2 + kt)^2}. \tag{4}$$

This quantity has been measured experimentally as a function of time for certain polycondensation processes by H. Dostal,[66, 67] P. J. Flory [9] and R. Raff.[8] It agrees fairly well with the requirements of equation (4). In particular, by varying the initial concentration of the monomer or the monomers, it was confirmed that the second-order character of equations (1) and (2) is correct and that the absolute value of k for a given concentration and temperature is the same as for a normal esterification. TABLE 1 contains the activation energies of normal esterifica-

TABLE 1

ESTERIFICATION PROCESS	E
Aliphatic acids in ethanol[a]	15,000
Phenylbenzoate in ethanol and water[b]	16,500
Ethylene glycol + phthalic acid[c]	22,600
Succinic acid + butylene glycol[d]	15,000
Polyesterification of diethylene glycol + adipic acid with *p*-toluenesulfonic acid as catalyst[e]	11,150
Polyesterification of diethylene glycol + adipic acid with no solvent and no catalyst[f]	13,000
− oxyundecanoic acid[g]	11,800
Cresol + formaldehyde[h]	20,000
Decamethylene glycol + decamethylene adipate with *p*-toluenesulfonic acid as catalyst[i]	12,150
Normal amidification[j]	38,000

[a] Fairclough, R. A., & Hinshelwood, C. N. Jour. Chem. Soc. 7: 593. 1939.
[b] Webers, W. A. Jour. Am. Chem. Soc. 58: 1014. 1936.
[c] Kienle, R. H., & Hovey, A. G. Jour. Am. Chem. Soc. 61: 3636. 1939
[d] Dostal, H., & Raff, R. Mh. Chem. 68: 188. 1936.
[e] Flory, P. J. Jour. Am. Chem. Soc. 62: 2261. 1940
[f] Flory, P. J. Jour. Am. Chem. Soc. 61: 3334. 1939.
[g] Davis, M. M. Trans. Faraday Soc. 34: 410. 1938.
[h] Dostal, H., & Raff, R. Mh. Chem. 68: 188. 1936.
[i] Flory, P. J. Jour. Am. Chem. Soc. 62: 2261. 1940.
[j] E. Pascu. Jour. Franklin Inst. 230: 132. 1940.

tions and polyesterifications and shows that they agree fairly well. It seems, therefore, that the basic assumption of a universal value of k is supported by comparison of equation (4) with experiments.

A further check can be made by considering that $m_x \cdot x \cdot M_0'$ is the total weight of the material comprised in x-mers, and $M_x = m_x \cdot x$ is the corresponding weight fraction. Its dependence upon x represents the

differential weight distribution curve of the polymer. According to equation (3) we get

$$M_x = x\ p^{x-1}(1-p)^2, \tag{5}$$

where p is given as a function of time by equation (3a).

Flory has carried out fractionations of polyesters and has shown that the general shape of the weight distribution curve agrees with equation (5), showing again that the assumptions used to derive this expression are supported by experiment.

It seems permissible, therefore, to consider each single elementary step of a polyesterification as a normal ester formation, disregarding whether the reacting group is at the end of a short or a long chain. As a consequence, it is to be expected that in a mixture of such polycondensed chains, hydrolytic and alcoholytic elementary steps take place which tend to reach, under given conditions, a certain polycondensation equilibrium. Flory [12, 13] has discussed such equilibria and has shown that they can be verified by experiment.

Finally, it may be added that the ratio between the weight and number average molecular weight is represented by

$$\frac{\overline{M}_w}{\overline{M}_n} = 1 + p,$$

showing that as p approaches 1, the weight average becomes twice the number average.

Another type of polyreaction which can be characterized by a *single rate constant* is represented by successive addition of a monomer to a functional group or a radical. If one introduces a certain number of active nuclei or functional groups into a system the molecules of which can add to such active centers, polymerization takes place according to the scheme:

$$\begin{aligned} m_1 + m &\longrightarrow m_2, \\ m_2 + m &\longrightarrow m_3 \\ m_i + m &\longrightarrow m_{i+1}. \end{aligned}$$

m_1 represents the functional molecule or radical which starts the whole process. It can be originally put into the system in a certain constant amount, or it can be slowly formed during the reaction. The former is the case if one initiates an addition polymerization by a catalyst or by ultraviolet light; the latter occurs if a slow nucleus formation cooperates with a fast propagation reaction. Chalmers,[65] Dostal [66] and Flory [11] have developed formulas which express the result of such a process. Its characteristic distinction from a polycondensation is that *only*

monomer addition takes place, but that *interaction of polymer with polymer* does not occur. It differs from the typical radical chain processes by the fact that there is no *termination or cessation* involved in it.

The addition of ethylene oxide to glycols, amines or acids is a typical case of such an *addition polymerization*. The initiating functional group can be OH, NH_2, NH, SH or COOH, the propagating group is the OH group, provided by the addition of the oxide.

Let N_0 be the number of ethylene oxide molecules and N_1, N_2, N_3 be the numbers of the species m_1, m_2, m_3 $\cdot$ m_x, having zero, one, two, three, etc., added ethylene oxide molecules. Then, the disappearance of the initiating material, m_1, will be given by

$$\frac{dN_1}{dt} = -kN_1. \tag{6}$$

The rate constant, k, includes the ethylene oxide concentration, which is so large that it does not change appreciably during the reaction. Considering N_2, it is apparent that it is produced at the same rate as N_1 disappears, according to equation (6), and is consumed according to the rate, $-kN_2$, hence

$$\frac{dN_2}{dt} = kN_1 - kN_2 \tag{7}$$

and

$$\frac{dN_x}{dt} = kN_{x-1} - kN_x. \tag{8}$$

The solution of this set of equations is

$$\frac{N_x}{N_0} = \frac{e^{-v}v^{x-1}}{(x-1)!}, \tag{9}$$

where v is defined by

$$dv = k\,dt.$$

The distribution (9) is considerably sharper than the one for polycondensation products. It has been discussed particularly by Dostal [66] and Flory,[11–13] but there are no experiments available as yet to check whether or not it is correct.

The ratio of the weight and number average molecular weight is given by

$$\frac{\bar{M}_w}{\bar{M}_n} = 1 + \frac{v}{(1+v)^2} \tag{10}$$

This shows that with increasing k and increasing time, the ratio $\bar{M}_w/\bar{M}_n$ approaches unity, which represents a comparatively homogene-

ous material whose osmotic and viscometric molecular weights should be very close together H Hibbert, R Fordyce and E L Lovell [5] have prepared very homogeneous polyoxyethylenes and studied their viscosities, but no experiments exist at present which would allow a direct check of equation (10)

Polycondensation and polymerization of the type just described are the only reactions which can be characterized by *one single rate constant* The larger majority of polymerizations seems to be of a much more complicated character We shall now pass to their description

POLYREACTIONS CONSISTING OF INITIATION, PROPAGATION AND TERMINATION

Many polymerization reactions particularly those which involve vinyl-, acrylic acid- and butadiene derivatives seem to be constituted mainly by three distinct elementary steps, which in some more complicated cases are joined by a few others (branching chain transfer, etc)

There are (compare TABLE 2) first *activation processes*, which convert

TABLE 2

(a) ACTIVATION PROCESSES

Nature of Process	$m_1 \to m_1^*$	$m_1 + X \to Xm$	$m_1 + m_1 \to m^*$
First Order Monomolecular	$k_{11}^1 c_1$		
Collision with Solvent		$k_{11}^2 c_1 c_l$	
Collision with Catalyst		$k_{11}^3 c_1 c_k$	
Proportional Catalyst Activity		$k_{11}^4 c_k$	
Photochemically		$k_{11}^5 c I$	
Bimolecular Second Order			$k_{12} c_1^2$

(b) PROPAGATION PROCESSES

Nature of Process	$m_j^* + m_1 \to m^*_{j+1}$	$m_j^* + m_1 \to m_j + m_1^*$
Normal Chain Growth	$k_2^1 c_j^* c_1$	
Chain Transfer		$k_2^2 c_j^* c_1$

(c) TERMINATION PROCESSES

Nature of Process	$m_j^* \to m_j$	$m_j^* + m \to m_{j+1}$	$m_j^* + s \to m_j + s$	$m_j^* + m_k^* \to m_{j+k}$
Monomolecular Isomerization	$k_{31} c_j^*$			
Collision with Monomer		$k_{32}^1 c_j^* c_1$		
Collision with Solvent			$k_{32}^2 c_j^* c_s$	
Collision between two Radicals				$k_{32}^3 c_j^{*2}$

the stable, unreactive monomer, m_1, into an activated molecule, m_1^*, m_2^* or km_1^* having the capacity of adding further monomers or other molecules. In all such cases the left-hand side of the chemical equation, which represents this step, contains no reactant with an asterisk (indicating activation), while there is always one (or two) asterisk on the right hand side of the chemical equation. *Activation* is *produced* by these processes.

Then, there are *propagation steps*, during which monomer is consumed and polymer built up, while *activation* is *maintained*. There is an asterisk on each side of the chemical equation, representing elementary steps of this kind.

Finally, there are *termination processes*, which *destroy activation* and hence have an asterisk only on the left-hand side.

The first part of TABLE 2 enumerates some *activation processes* which have been found to occur experimentally. The monomer is (as usual) represented by m_1; its concentration (or better its activity) by c_1. An activated monomer, m_1^*, can be produced by a first-order reaction either monomolecularly or by a collision with a solvent molecule (activity c_s). An activated dimer, m_2^*, can be the result of a sufficiently vigorous collision between two monomers. Activated monomers can be produced by collision with a catalyst (activity c_k), a solvent (c_s), or by absorption of a light quantum (intensity I). There may be other types of activation (termolecular or by collision of the monomer, m_1, with a polymer, m_j), but those listed in TABLE 2a seem to be most frequent.

TABLE 2b contains *propagation steps*. First, the bimolecular, second-order addition of m_1 to an activated polymer m_j^*, which seems to represent normal chain growth. The process in the next line consists in the transfer of the activation (asterisk) from a chain with j links, m_j^*, to a monomer, m_1^*. This process was first considered by Flory [10] and is usually termed *chain transfer*. It does not destroy the growth potential as such, but shifts the polymerization degree to a lower average value.

TABLE (2c) finally lists some of the more important *cessation reactions*. The first line represents a monomolecular, first-order termination, which can be ring formation or isomerization of a biradical. The second line lists termination by collision with an inactive monomer, which may consist in the exchange of two H-atoms between the monomer and a biradical. The next line shows the same thing for collision with a solvent molecule and the last contains the mutual saturation of two activated chains (each of which can have any degree of polymerization, including the monomer). This type of termination seems to be preponderant,

if the growing chains have one unpaired electron at one of their ends (case $k_{11}{}^4$ of 2a) and exhibit the nature of a normal organic radical.

The expressions of TABLE 2 permit a simple classification of polyreactions of this type, which is, in a certain sense, similar to the division of ordinary reactions into mono-, bi- and trimolecular ones. According to them, active centers, c^* (regardless of chain length), are produced by a certain process, say by

$$+\frac{dc^*}{dt} = k_{12}c_1{}^2 \tag{11}$$

and consumed by another process, say by

$$-\frac{dc^*}{dt} = k_{32}c^{*2}. \tag{12}$$

At the very beginning of the experiment, c^* is zero and hence there will be only production of active centers according to reaction (11), but no consumption according to reaction (12). However, as soon as a certain concentration of c^* is built up by reaction (11), $k_{32}c^{*2}$ will gradually increase until $-\frac{dc^*}{dt}$ becomes equal to $+\frac{dc^*}{dt}$, from which moment a steady value of c^* is reached, which remains constant as long as there is enough monomer left to maintain it according to reaction (11). This steady state consideration was first introduced by Bodenstein and is a very useful approximation to deal with chain reactions of different kinds. Recently, R. Ginell and R. Simha [14] investigated under what conditions the steady state method can be legitimately applied in polymerization reactions and developed equations which hold in any case.

Setting the left sides of (11) and (12) equal, we obtain for the steady state concentration of active (growing) centers in the systems

$$c^* = \sqrt{\frac{k_{12}}{k_{32}}}\, c_1. \tag{13}$$

We can now combine all individual cases of TABLE 2a and 2c with each other and obtain the corresponding steady state concentrations as listed in TABLE 3.

It may be possible under certain favorable conditions, to carry out a direct measurement of c^* (using an inhibitor or the magnetic method). However, it is possible to compute with the aid of c^* other quantities which can be observed and measured directly.

TABLE 3

STEADY STATE CONCENTRATION OF ACTIVE (GROWING) CENTERS UNDER VARIOUS ASSUMPTIONS REGARDING INITIATION AND TERMINATION

Termination \ Initiation	$k_{11}^1 c_1$	$k_{11}^2 c_1 c_s$	$k_{11}^3 c_1 c_k$	$k_{11}^4 c_k$	$k_{11}^5 c_1 I$	$k_{12} c_1^2$
$k_{31} c^*$	$\frac{k_{11}^1}{k_{31}} c_1$	$\frac{k_{11}^2}{k_{31}} c_1 c_s$	$\frac{k_{11}^3}{k_{31}} c_1 c_k$	$\frac{k_{11}^4}{k_{31}} \cdot c_k$	$\frac{k_{11}^5}{k_{31}} c_1 I$	$\frac{k_{12}}{k_{31}} \cdot c_1^2$
$k_{32}^1 c^* c_1$	$\frac{k_{11}^1}{k_{32}^1}$	$\frac{k_{11}^2}{k_{32}^1} c_s$	$\frac{k_{11}^3}{k_{32}^1} c_k$	$\frac{k_{11}^4}{k_{32}^1} \frac{c_k}{c_1}$	$\frac{k_{11}^5}{k_{32}^1} I$	$\frac{k_{12}}{k_{32}^1} \cdot c_1$
$k_{32}^2 c^* c_s$	$\frac{k_{11}^1}{k_{32}^2} \frac{c_1}{c_s}$	$\frac{k_{11}^2}{k_{32}^2} c_1$	$\frac{k_{11}^3}{k_{32}^2} \frac{c_1 c_k}{c_s}$	$\frac{k_{11}^4}{k_{32}^2} \frac{c_k}{c_s}$	$\frac{k_{11}^5}{k_{32}^2} \frac{c_1 I}{c_s}$	$\frac{k_{12}}{k_{32}^2} \cdot \frac{c_1^2}{c_s}$
$k_{32}^3 c^{*2}$	$\left(\frac{k_{11}^1}{k_{32}^3}\right)^{1/2} c_1^{1/2}$	$\left(\frac{k_{11}^2}{k_{32}^2}\right)^{1/2} (c_1 c_s)^{1/2}$	$\left(\frac{k_{11}^3}{k_{32}^2}\right)^{1/2} (c_1 c_k)^{1/2}$	$\left(\frac{k_{11}^4}{k_{32}^3}\right)^{1/2} (c_k)^{1/2}$	$\left(\frac{k_{11}^5}{k_{32}^3}\right)^{1/2} (c_1 I)^{1/2}$	$\left(\frac{k_{12}}{k_{32}^3}\right)^{1/2} \cdot c_1$

First, it is to be remembered that according to TABLE 2b the monomer is consumed at a rate of

$$-\frac{dc_1}{dt} = (k_{22}^1 + k_{22}^2)c^*c_1, \tag{14}$$

k_{22}^1 being the rate constant of normal propagation, k_{22}^2, the one of chain transfer. There is, according to TABLE 2c, another monomer consumption involved in $m_j^* + m_1 \rightarrow m_{j+1}$, so that we obtain for the disappearance of the monomer (over-all rate of the reaction)

$$-\frac{dc_1}{dt} = (k_{22}^1 + k_{22}^2 + k_{32}^1)c^*c_1 = k_2c^*c_1. \tag{15}$$

Work in this direction has just only begun.[71a]

Introducing here the values for c^* as listed in TABLE 3, we can easily write the rate with which the monomer disappears during the polymerization. This has been done in TABLE 4. It may be pointed out that the most reliable values for $-\frac{dc_1}{dt}$ are obtained at the beginning of the polymerization (after any eventual induction period has disappeared), where the monomer is still in excess and no appreciable amount of polymer has yet been formed.

The expressions of TABLE 4 can be directly compared with the rate of (initial) monomer consumption measured as a function of the activities of monomer, solvent, catalyst, intensity of light, etc. They contain the individual rate constants of all elementary steps involved. As we have pointed out before, our final goal is to determine the absolute values and temperature dependence (A and E values) of all these individual rate constants, in order to use them for the elucidation of their chemical nature. By measuring the over-all rate of the polymerization we can, according to TABLE 4, only determine a *combination* of these individual rate constants but not each of them individually. In order to do the latter, we need additional measurements and additional equations to evaluate and to coordinate them with the results of TABLE 4.

One additional item of information can be provided by the number average of the degree of polymerization, $\bar{P}_n$, corresponding to the polymer formed at the beginning of the reaction. On the average, each active nucleus will have the same chance to grow out into a chain before it is terminated and therefore the same number of monomers (number average for degree of polymerization $\bar{P}_n$) will correspond to each nucleus. Hence,

71a **Goldfinger, G., Skeist, I., & Mark, H.** Paper presented at the Acad. Sci. meeting in Pittsburgh, September, 1943.

TABLE 4

INITIAL OVER-ALL RATE OF THE REACTION UNDER VARIOUS ASSUMPTIONS REGARDING INITIATION, PROPAGATION AND TERMINATION

Termination \ Initiation	$k_{11}^1 c_1$	$k_{11}^2 c_1 c_s$	$k_{11}^3 c_1 c_k$	$k_{11}^4 c_k$	$k_{11}^5 c_1 I$	$k_{12} c_1^2$
$k_{31} c^*$	$\frac{k_{11}^1 k_2}{k_{31}} c_1^2$	$\frac{k_{11}^2 k_2}{k_{31}} c_1 c_s$	$\frac{k_{11}^3 k_2}{k_{31}} c_1^2 c_k$	$\frac{k_{11}^4 k_2}{k_{31}} c_1 c_k$	$\frac{k_{11}^5 k_2}{k_{31}} c_1^2 I$	$\frac{k_{12} k_2}{k_{31}} \cdot c_1^3$
$k_{32}^1 c^* c_1$	$\frac{k_{11}^1 k_2}{k_{32}^1} c_1$	$\frac{k_{11}^2 k_2}{k_{32}^1} c_1 c_s$	$\frac{k_{11}^3 k_2}{k_{32}^1} c_1 c_k$	$\frac{k_{11}^4 k_2}{k_{32}^1} c_k$	$\frac{k_{11}^5 k_2}{k_{32}^1} c_1 I$	$\frac{k_{12} k_2}{k_{32}^1} \cdot c_1^2$
$k_{32}^2 c^* c_s$	$\frac{k_{11}^1 k_2}{k_{32}^2} \frac{c_1^2}{c_s}$	$\frac{k_{11}^2 k_2}{k_{32}^2} c_1^2$	$\frac{k_{11}^3 k_2}{k_{32}^2} \frac{c_1^2 c_k}{c_s}$	$\frac{k_{11}^4 k_2}{k_{32}^2} \frac{c_1 c_k}{c_s}$	$\frac{k_{11}^5 k_2}{k_{32}^2} \frac{c_1^2 I}{c_s}$	$\frac{k_{12} k_2}{k_{32}^2} \cdot \frac{c_1^3}{c_s}$
$k_{32}^3 c^{*2}$	$\left(\frac{k_{11}^1}{k_{32}^3}\right)^{1/2} k_2 c_1^{3/2}$	$\left(\frac{k_{11}^2}{k_{32}^3}\right)^{1/2} k_2 c_1^{3/2} c_s^{1/2}$	$\left(\frac{k_{11}^3}{k_{32}^3}\right)^{1/2} k_2 c_1^{3/2} c_k^{1/2}$	$\left(\frac{k_{11}^4}{k_{32}^3}\right)^{1/2} k_2 c_1 c_k^{1/2}$	$\left(\frac{k_{11}^5}{k_{32}^3}\right)^{1/2} k_2 c_1^{3/2} I^{1/2}$	$\left(\frac{k_{12}}{k_{32}^3}\right)^{1/2} k_2 c_1^2$

the rate of monomer consumption, $\frac{dc_1}{dt}$, will be obtained by multiplying the rate of the activation by the average number of monomer molecules bound in one chain

$$\frac{dc_1}{dt} = \frac{dc^*}{dt} \cdot \bar{P}_n. \tag{16}$$

This, however, is true only if chain transfer is negligible compared with propagation.

Using the expressions of TABLES 2 and 4, we obtain TABLE 5, which shows how the number average for the degree of polymerization of the initially produced polymer depends upon the individual rate constants involved. As already indicated above, $\bar{P}_n$ can be measured experimentally and we have therefore a second equation for the determination of the individual rate constants.

Let us consider a few examples

(a) Assume that the nuclei are formed by pure thermal collisions according to $k_{12}c_1^2$, and that the chains are propagated according to $k_{22}^1c^*c_1$ and terminated by $k_{32}^3c^{*2}$. Then we obtain from TABLES 4 and 5:

$$\text{(A)}\left\{\begin{array}{l} \text{Initial rate of monomer consumption} = k_{22}^1\left(\dfrac{k_{12}}{k_3{}^1}\right)^{1/2} c_1^2 \text{ and} \\[2ex] \text{Number average degree of polymerization} = \dfrac{k_{22}^1}{(k_{12}k_{32}^3)^{1/2}} \end{array}\right.$$

Dividing these two expressions, we must expect to get the rate of nucleus formation and, in fact, we obtain $k_{12}c_1^2$, as assumed above.

In this case we should expect a second-order over-all rate and independence of the degree of polymerization upon monomer concentration. It seems that this is true in certain cases of high temperature ethylene and styrene polymerization in the gas phase,[8] but it does not seem to present the ordinary picture of vinyl-polymerization.

(b) Assume that nuclei are formed by collision between the monomer and a catalyst, $k_{11}^3c_1c_k$, built up into chains by $k_{22}^1c^*c_1$ and destroyed by $k_{32}^3c^{*2}$. Then we obtain

$$\text{(B)}\left\{\begin{array}{l} \text{Initial over-all rate of polymerization} = k_{22}^1\left(\dfrac{k_{11}^3}{k_{32}^3}\right)^{1/2} c_1^{3/2}c_k^{1/2} \quad \text{and} \\[2ex] \text{Number average degree of polymerization} = \dfrac{k_{22}^1}{(k_{11}^3k_{32}^3)^{1/2}}\left(\dfrac{c_1}{c_k}\right)^{1/2} \end{array}\right.$$

TABLE 5

NUMBER AVERAGE POLYMERIZATION DEGREE OF THE POLYMER FORMED AT THE BEGINNING OF THE STEADY STATE PERIOD USING VARIOUS ASSUMPTIONS REGARDING INITIATION, PROPAGATION AND TERMINATION

Termination \ Initiation	$k_{11}^1 c_1$	$k_{11}^2 c_1 c_s$	$k_{11}^3 c_1 c_k$	$k_{11}^4 c_k$	$k_{11}^5 c_1 I$	$k_{12} c_1^2$
$k_{31} c^*$	$\frac{k_{22}^1}{k_{31}} c_1$	$\frac{k_{22}^1}{k_{31}} c_1$	$\frac{k_{22}^1}{k_{31}} c_1$	$\frac{k_{22}^1}{k_{31}} c_1$	$\frac{k_{22}^1}{k_{31}} c_1$	$\frac{k_{22}^1}{k_{31}} c_1$
$k_{32}^1 c^* c_1$	$\frac{k_{22}^1}{k_{32}^1}$	$\frac{k_{22}^1}{k_{32}^1}$	$\frac{k_{22}^1}{k_{32}^1}$	$\frac{k_{22}^1}{k_{32}^1}$	$\frac{k_{22}^1}{k_{32}^1}$	$\frac{k_{22}^1}{k_{32}^1}$
$k_{32}^2 c^* c_s$	$\frac{k_{22}^1}{k_{32}^2} \frac{c_1}{c_s}$	$\frac{k_{22}^1}{k_{32}^2} \frac{c_1}{c_s}$	$\frac{k_{22}^1}{k_{32}^2} \frac{c_1}{c_s}$	$\frac{k_{22}^1}{k_{32}^2} \frac{c_1}{c_s}$	$\frac{k_{22}^1}{k_{32}^2} \frac{c_1}{c_s}$	$\frac{k_{22}^1}{k_{32}^2} \cdot \frac{c_1}{c_s}$
$k_{32}^3 c^{*2}$	$\frac{k_{22}^1}{(k_{11}^1 k_{32}^3)^{1/2}} \cdot c_1^{1/2}$	$\frac{k_{22}^1}{(k_{11}^2 k_{32}^3)^{1/2}} \left(\frac{c_1}{c_s}\right)^{1/2}$	$\frac{k_{22}^1}{(k_{11}^3 k_{32}^3)^{1/2}} \left(\frac{c_1}{c_k}\right)^{1/2}$	$\frac{k_{22}^1}{(k_{11}^4 k_{32}^3)^{1/2}} \frac{c_1}{c_k^{1/2}}$	$\frac{k_{22}^1}{(k_{11}^5 k_{32}^3)^{1/2}} \left(\frac{c_1}{I}\right)^{1/2}$	$\frac{k_{22}^1}{(k_{12} k_{32}^3)^{1/2}}$

These two expressions show that àddition of a catalyst speeds up beneficially the over-all rate of polymerization. It reduces, however, in a very undesired way the degree of polymerization. It has been found by various investigators that the over-all rate of vinyl type polymerization proceeds with the square root of the catalyst concentration [18, 20, 27, 34, 37, 59, 72] and in many cases this fact has been interpreted as proof for a dissociation of the catalyst into radicals, before it interferes with the monomer. Although this is a probable interpretation and is supported by other experimental evidence,[41, 72-76] it is worthwhile to emphasize that, from the point of view of formal kinetics, this square-root law is only a consequence of first-order nucleus formation with respect to the monomer and to the catalyst, and second-order termination with respect to the active centers.

Over-all reaction rates proportional to a low power of the monomer concentration have been observed on several occasions and it seems that case (B) comes near to the conditions of peroxide catalyzed styrene polymerization in high monomer concentration.

(c) Let us now assume, for a slight change of the conditions of case (B), that the monomer is in such excess and the catalyst concentration so low that nucleus formation depends on the catalyst concentration alone, according to $k_{11}^4 c_k$, and chain propagation is controlled by $k_{22}^1 c^* c_1$ and termination by $k_{32}^3 c^{*2}$. Then we get

$$
\text{(C)}\begin{cases} \text{Initial over-all rate of the reaction} = k_{22}^1 \left(\dfrac{k_{11}^4}{k_{32}^3}\right)^{1/2} c_1 c_k^{1/2} \text{ and} \\[2ex] \text{Average degree of polymerization} = \dfrac{k_{22}^1}{(k_{11}^4\ k_{32}^3)^{1/2}}\,\dfrac{1}{c_k^{1/2}}. \end{cases}
$$

This is the case which was recently investigated very thoroughly by C. C. Price and his collaborators.[34, 39-41] Since it will be fully explained and discussed in Professor Price's paper, it may be appropriate not to anticipate too much, but to refer to his article in this publication

(d) Let us finally ask ourselves which kind of catalyzed polymerization would be most advantageous from the technical point of view, using the catalyst only for what it should be—namely, as an accelerator for the over-all rate of the polymerization, without having it interfere unbeneficially with the degree of polymerization.

Looking at our tables, we find immediately that the most favorable

---

[72] **Cuthberson, A. C., Gee, G., & Rideal, E. K.** Nature **140:** 889. 1937

[73] **Breitenbach, J. W., & Maschin, H.** Zeit. physikal Chem **A187:** 175 1940

[74] **Hey, M. H., & Waters, V. F.** Chem. Rev. **21:** 169. 1937.

[75] **Houtz, R. C., & Adkins, H.** Jour. Am. Chem. Soc **53:** 1058. 1931, **55:** 1609 1933.

[76] **Ziegler, K.** Chem Ztg **62:** 125. 1938.

conditions would be obtained if nucleus formation occurred according to $k_{11}{}^{3}c_1c_k$, chain growth according to $k_{22}{}^{1}c_1c^*$ and termination by $k_{31}c^*$. In this case we get

$$
\text{(D)}\begin{cases}
\text{Over-all rate} = \dfrac{k_{11}{}^{3}k_{22}{}^{1}}{k_{31}}\cdot c_1{}^{2}c_k \\[2ex]
\text{Degree of polymerization} = \dfrac{k_{22}{}^{1}}{k_{31}}\, c_1.
\end{cases}
$$

Increase of monomer concentration speeds up the reaction with the square of the concentration and increases the molecular weight linearly, while catalyst addition accelerates the over-all rate linearly, without affecting the degree of polymerization. This double advantage is brought about by the first-order termination according to $k_{31}c^*$. Recent experiments by J. Abere, G. Goldfinger and H. Naidus [20] and by R. G. W. Norrish and R. R. Smith [77] suggest that such conditions can be realized, if the polymer grows in a heterogeneous system in which a gel phase and a solution phase co-exist. It may also be that emulsion and pearl polymerization offer analogous advantages. Unfortunately, with the exception of a number of patent claims, there is no quantitative experimental material available at present which would permit one to follow more thoroughly this line of approach.

It seems, however, that the systematic treatment as initiated by various authors [3, 14, 17, 31, 65, 66, 78] and as adopted and recommended in this article, might have a certain heuristic value in order to establish favorable conditions for steering and controlling polymerization reactions.

Nevertheless, it must be remembered that this comprehensive treatment represents only a first approximation, inasmuch as the steady state principle is used and no complete integration of the general differential equations is carried out. Hence, we only get information as to the *over-all rate* and as to the *average degree of polymerization*, but do not obtain expressions for the molecular size distribution curve. It must also be emphasized that the two quantities, $\dfrac{dc_1}{dt}$ and P_n, which are given in TABLES 4 and 5 and which can be compared directly with experimental data, are not sufficient to determine unambiguously the individual rate constants of the different elementary steps. We have two equations, but at least three (in many cases more) unknown quantities. Hence, we

[77] Norrish, R. G. W., & Smith, R. R. Nature 150: 336. 1942.
[78] Hulburt, H., Harman, R. A., Tobolsky, A., & Eyring, H. Ann. N. Y. Acad. Sci. 44: 371. 1943.

have to look for further expressions which combine these same unknowns (k_{11}^2, k_{22}^1, k_{32}^2, etc.) with other directly measurable quantities. One of these is the distribution curve of the polymer formed after different lengths of time and under various experimental conditions. If it would be possible to obtain theoretical expressions for the molecular size distribution as a function of time and of the various individual rate constants, and if simultaneously the distribution curve could be determined experimentally, we could compute the rate constants of the different reaction steps individually.

Very promising attempts have been made in this direction recently by R. Ginell and R. Simha [14] and by H. Eyring, R. A. Harman, H. M. Hulburt and A. V. Tobolsky.[78] The investigations of Eyring and his collaborators are presented in another article in this publication. It will be sufficient here to draw special attention to their very interesting results and to present and discuss briefly the results arrived at by Ginell and Simha.[14]

These authors have succeeded in carrying through the integration of the differential equations under rather general conditions They consider activation of monomolecular and bimolecular character

$$m_1 \xrightarrow{k_{11}} m_1^*, \tag{17a}$$

$$m_1 + m_1 \xrightarrow{k_{12}} m_1^* + m_1. \tag{17b}$$

The first index in the rate constants refers to the elementary step (1 = activation, 2 = propagation, 3 = termination); the second, to the order of the elementary reaction involved.

The propagation is described by

$$m_j^* + m_1 \xrightarrow{k_{22}} m_{j+1}^* \tag{18}$$

and the cessation by

$$m_j^* + m_1 \xrightarrow{k_{32}} m_{j+1}. \tag{19}$$

Using the elementary steps as represented by equations (17), (18) and (19), the differential equations are:

$$\frac{dm_1}{dt} = -k_{11}m_1 - k_{22}m_1 \sum_1^\infty m_j^* - k_{32}m_1 \sum_1^\infty m_j^*, \tag{20a}$$

$$\frac{dm_1^*}{dt} = +k_{11}m_1 - k_{22}m_1m_1^* - k_{32}m_1m_1^*; \tag{20b}$$

$$\frac{dm_j^*}{dt} + k_{22}m_1m_{j-1}^* - k_{22}m_1m_j^* - k_{32}m_1m_j^*, \tag{20c}$$

$$\frac{dm_j}{dt} + k_{32}m_1m_{j-1}^*. \tag{20d}$$

Equation (20a) lists all possibilities for the consumption of the inactivated monomer, m_1; they are:

(a) its conversion into an activated molecule by an initiation reaction of the first order according to $k_{11}m_1$,

(b) its building into growing chains of all lengths by a second-order reaction, and

(c) its consumption during a termination process of the second order.

Equation (20b) establishes the balance for the activated monomer. It is produced by the initiation process according to $k_{11}m_1$, and is used up by two reactions:

(a) the propagation according to $k_{22}m_1m_1^*$, whereby it is converted into m_2^*, and

(b) the termination, whereby it is transformed into m_2 according to $k_{32}m_1m_1^*$.

Equation (20c) considers production and consumption of chain molecules having the arbitrary degree of polymerization, j. They are formed by chain growth from molecules m_{j-1} and are converted

(a) by propagation, into m_{j+1}^* according to $k_{22}m_1m_j^*$ or

(b) by termination, into m_{j+1} according to $k_{32}m_1m_1^*$.

Equation (20d) finally accounts for the accumulation of the stable molecules of the general type m_j by termination of chains m_{j-1}^* through a collision with m_1.

It is apparent that the system (20) is fairly flexible and allows accounting for all kinds of other assumptions concerning the different elementary steps. If, for example, the initiation is of the second order in m_1, then the first terms in equation (20a) have to be changed to $-k_{12}m_1^2$ and $+k_{12}m_1^2$, respectively. If a monomolecular deactivation of m_1^* takes place, one has to add a term, $-k_{31}m_1^*$ in (20a). Presumably there will be little change necessary in the propagation terms—terms 2 in (20a, b) and term 1 in (20c), because chain growth will always be of the second order in m_1 and m_j^*. Several possible changes suggest themselves for the cessation terms (last terms in all equations). There may be chain breaking by the solvent or by any impurity present in the system, which can be expressed by a termination according to $k_{31}m_j^*$, where k_{31} contains the concentration or activity of the chain-breaking substance. Mutual termination by collision of two growing chains

would lead to a term, $k_{32}m_i{}^*m_j{}^*$, in (20b) and to corresponding terms in the other equations.

Ginell and Simha [14] have carried out the integration of equations (20) for the cases of first- and second-order initiation and have worked out a set of formulas and curves which describe the course of the reaction in all its details. It would be outside of the scope of this comprehensive article to give a complete description of their article, but it seems appropriate to mention at least its main results.[79]

The final number distribution function of the polymer is given by

$$m_j = m_1{}^0 \left(\frac{k_{32}}{k_{22}}\right)^2 \left(\frac{k_{22}}{k_{22} + k_{32}}\right)^{j-1}, \tag{21}$$

$m_1{}^0$ being the amount of monomer originally present and, accordingly, the final weight fraction distribution function is

$$w_j = m_1{}^0\, j \left(\frac{k_{32}}{k_{22}}\right)^2 \left(\frac{k_{22}}{k_{22} + k_{32}}\right)^{j-1} \tag{22}$$

Equation (22) gives us exactly what we were looking for. It contains the rate constants of *propagation* and *termination* and combines them with a new experimentally accessible quantity, namely the *weight fraction*, w_j.

No systematic and entirely satisfactory experimental investigation has yet been published which utilizes the combination of equations (15), (16) and (22), but recently T. Alfrey [19] and A. Bartovics [4] have worked out final weight-fraction distribution curves of three polystyrene samples which had been prepared with benzoyl peroxide at 60°, 120° and 180° C and determined osmotic and viscometric molecular weights of the various fractions. Combining their results with the findings of J. Abere, G. Goldfinger and H. Naidus [20] and with previous studies of Breitenbach,[29–31] Melville,[53] Norrish,[62] Pfann,[37] Raff [3] and Schulz,[16] it seems that one can represent the course of this reaction with fair approximation by three rate constants $k_{11}{}^4$, $k_{22}{}^1$ and $k_{32}{}^3$, which are characterized by the following frequency factors and activation energies:

$k_{11}{}^4$	A about 10^{11}; E = 24,000 cal. per mol.
$k_{22}{}^1$	A about 10^5; E = 5,000 cal. per mol.
$k_{32}{}^3$	A about 10^4; E = 5,000 cal. per mol.

[79] The authors are very much indebted to Mr R Ginell and Dr. R. Simha for having discussed these results with them before they were published

The absolute values of these rate constants at 120–130° C. are:

rate of initiation	about 10^{-2}
rate of propagation	about 10^{2}
rate of termination	about 10^{1}

and their ratio is:

propagation	10^{4} times faster than initiation, and
propagation	10 times faster than termination.

From TABLE 5 we get for the average degree of polymerization about 4000, which is on the high side, presumably because chain transfer and branching have not been taken into account.

The above figures make it legitimate, according to Ginell and Simha, to apply the Bodenstein principle of a steady state of the active centers and justify *a posteriori* the use of TABLES 3, 4 and 5 in the discussion of this kind of vinyl type polymerizations. It must be pointed out, however, that there are indications that at higher temperatures chain transfer and branching interfere with the results and have to be more properly accounted for.

This brief description of the formal kinetics of vinyl type polymerization may serve to show that, up to date, we have only to a very limited extent approached the goal of a quantitative resolution of the course of such a reaction in its various individual steps. Keeping this preliminary state of our knowledge in mind, we shall now pass to the description of the chemical nature of the different elementary steps.

THE ELEMENTARY STEPS OF VINYL TYPE POLYMERIZATION

At least three individual elementary steps have to be considered: *activation*, *propagation* and *termination*, but it seems that in many cases (if not in all) other processes, such as *chain transfer* and *branching* render the conditions more complicated, as represented in TABLES 3, 4 and 5. The two articles by C. C. Price and H. M. Hulburt, R. A. Harman, A. V. Tobolsky and H. Eyring give a very excellent and complete description of our present knowledge concerning the chemical nature of the different elementary steps. It may, therefore, be appropriate not to overlap these presentations, but to add here only a very short and cursory report, without going into any detailed discussion.

Activation Processes

It seems that all activation processes, as listed in TABLE 2, can and do take place under appropriate experimental conditions. The thermal (uncatalyzed) activation (first or second order) becomes noticeable at higher temperatures (styrene above 180°; vinyl acetate above 80° C.), while the catalyzed activation prevails at lower temperatures. In certain cases (free radicals, alkali metals, peroxides, ozonides) the catalyst acts by producing a free radical, in other instances (BF_3, $AlCl_3$, $SnCl_4$) association of the catalyst with the monomer takes place. Photopolymerization takes place mostly (or always) according to $\sqrt{I}$, showing that the photoactivated monomer can either return to the unactive state or grow into a chain.

Burk,[46] Breitenbach,[29,30] Goldfinger,[30] Lankelma,[46] Moore,[46] Raff,[3] Schulz [17] and Whitby [44] have investigated the uncatalyzed polymerization of styrene; it seems that an energy of about 23,000–25,000 cal. per mol. is necessary to produce activation and that the collision factor is between 10^4 and 10^6. Activation of this kind is a rather slow reaction at temperatures below 150° C. (k_{11}^1 or k_{12} about 10^{-8}).

Radical catalyzed polymerization of vinyl- and acryl-derivatives has been studied by Abere,[20] Alyea,[27] Bolland,[28] Goldfinger,[20] Naidus,[20] Norrish,[77] Pfann,[37] Price,[41] Raff [3] and Schulz.[18] Three ways have been chosen to accumulate additional information as to whether or not a radical mechanism is responsible for the activation:

(a) The dependence of the initial monomer consumption upon the concentration of the catalyst. TABLE 6 enumerates cases in which the square root of catalyst concentration was proportional to the initial reaction rate.

(b) The appearance of the dissociated catalyst molecule at one or both ends of the chains of the final polymer. TABLE 7 shows cases in which the dissociation products of appropriately substituted catalysts have been located analytically at the chain ends of the polymer. The number average molecular weight, as calculated from this end-group determination, has been compared with other number or weight average determinations and is in reasonable agreement with them.

(c) In one case the concentration of the activating radicals was measured directly by the color of the solution and it was found that the initial rate of monomer consumption is proportional to this concentration. Number average molecular weights as determined

TABLE 6

INITIAL RATE IS PROPORTIONAL TO SQUARE ROOT OF CATALYST CONCENTRATION

Substance	Catalyst	Temperature °C.	E cal, mol	A	Authors
Styrene	benzoyl peroxide	60–120	25,000	10^{11}	Breitenbach
Styrene	benzoyl peroxide	135–150	23,000	10^{11}	Raff
Styrene	benzoyl peroxide	60–120	23,000	10^{11}	Schulz
Methyl methacrylate	benzoyl peroxide	25–60	—	—	Norrish, Smith
Methyl methacrylate	benzoyl peroxide				Alyea
Methyl methacrylate	p-NO_2 benzoyl peroxide	80	—	—	Gartland
Methyl methacrylate	p-Cl benzoyl peroxide				
Methyl methacrylate	p-toloyl peroxide				Graham
Styrene	p-Br-benzoyl peroxide	40–100	24,000	10^{10}–10^{11}	Pfann
d-s-butyl-chloro acrylate	benzoyl peroxide	26–68	15,200	—	Price
Styrene	benzoyl peroxide in various solvents	60–140	23,000 to 26,000	10^{10}–10^{11}	Abere, Goldfinger, Mark, Naidus

TABLE 7

CATALYST HAS BEEN FOUND AT THE CHAIN ENDS

Substance	Catalyst	DP from end group analysis		DP osmotic or viscometric	Author
Styrene	p-Br diazonium-hydroxide	30 for Br 17 for Br	.Br OH	22 (visc)	Price and Durham
Styrene	p-Br-benzoylperoxide	40 for Br 19 for Br	Br X	34 (osm)	Plann
Styrene	tetraphenyl-succinonitril	N has been found in the polymer			Schulz
Styrene	tetra-Cl-tetraphenyl-succinonitril	21 for R	R	17 (osm)	Raff

from different sources agreed within the limit of experimental errors.

This seems to indicate that under certain conditions (vinyl-derivatives, low temperatures, easily dissociating catalysts) the activation to radicals plays an important role and decidedly influences the kinetics of the reaction. It has been suggested [3] that in such cases the activation of the ethylene double bond consists in lifting the two electrons of the second kind in a state of identical spin, so that they occupy repulsive orbitals. Recently, Eyring and his collaborators [78, 80–82] have shown that the singlet-triplet transition in ethylene derivatives not only accounts for the radical catalyzed polymerization of such compounds, but also explains very satisfactorily a number of cis-trans isomerization processes and certain typical addition rules of the double bond.

If one adopts this point of view, catalysts of the peroxide, ozonide and radical type accelerate the initiation reaction not by decreasing the activation energy, but by facilitating the singlet-triplet transition by magnetic perturbation, which increases the frequency factor from 10^5 to about 10^{11}. As far as uncatalyzed polymerization has been observed (compare above), the activation was found to occur according to $A = 10^5$ and $E = 23{,}000$ cal. per mol., while radical type catalysis always leads to values of $A = 10^{11}$, $E = 23{,}000$ cal. per mol.

Polymerization catalyzed by polarization of the double bond has been studied by Frolich,[83] Gwyn Williams,[84] Norrish [62] and others.[85] It

[80] **Harman, R. A., & Eyring, H.** Jour. Chem. Phys. 10: 557. 1942.
[81] **Magee, J. L., Strand, W., & Eyring, H.** Jour Am. Chem. Soc. 63: 677. 1941.
[82] **Ri, T., & Eyring, H.** Jour. Chem. Phys. 8: 433. 1940.
[83] **Thomas, R. M., Sparks, W. J., Frolich, P. K., & Muller-Cunradi, M.** Jour. Am Chem Soc. 62: 276. 1940.
[84] **Williams, G.** Jour. Chem. Soc London 1938: 246, 1046, 1940. 775.
[85] **Wasserman, A.** Trans. Faraday Soc. 34: 128. 1938.

seems to occur at low temperatures in the presence of strongly polar molecules, such as BF_3, $AlCl_3$, etc. Eyring[78] has pointed out that in such cases presumably the heteropolar elevated state of the double bond[86–89] is responsible for the activation. One should expect that in these instances the initial rate of monomer consumption is directly proportional to the concentration of the catalyst. It seems that this is true for the polymerization of isobutylene with BF_3 and of styrene with $SnCl_4$, although it would be very welcome to have better experimental evidence of this fact.

This type of catalysis accelerates the reaction by decreasing the activation energy, which is necessary to bring the molecule into the first activated singlet state and which is about 40,000 cal. per mol., down to values between 20,000 and 25,000 cal. per mol. The collision factor is about 10^{11} and is not noticeably affected by the catalyst.

Altogether, it appears that the two possible uncatalyzed initial reactions—the nonadiabatic (singlet-triplet) radical type activation and the adiabatic (singlet-singlet) ionic type activation—are both slow at temperatures below 200° C.; the former, because the frequency constant is small (about 10^5); the latter, because the activation energy is large (about 40,000 cal. per mol.). In both cases catalysis can be effectively applied; in the former, the frequency constant is increased up to about 10^{11}, in the latter, the activation energy is lowered down to about 20,000 cal. per mol. In both ways one arrives at reactions having a reasonable rate, which conforms fairly well with our experimental knowledge. TABLE 8 shows a few figures which may help to elucidate this situation.

TABLE 8

APPROXIMATE RATES OF UNCATALYZED AND CATALYZED VINYL TYPE POLYMERIZATION

Type of Reaction	A in sec.$^{-1}$	E in cal./mol.	Absolute Rate Constant At		
			50° C.	150° C.	250° C.
Uncatalyzed non-adiabatic (radical) type	10^5	24,000	10^{-11}	10^{-7}	10^{-5}
Catalyzed (peroxide) non-adiabatic type	10^{11}	24,000	10^{-5}	10^{-1}	10^{+1}
Uncatalyzed adiabatic (ionic) type	10^{11}	40,000	10^{-16}	10^{-9}	10^{-5}
Catalyzed (BF_3) adiabatic type	10^{11}	24,000	10^{-5}	10^{-1}	10^{+1}

[86] Huckel, E. Zeit. Phys. 60: 423. 1930
[87] Mulliken, R. S. Phys. Rev. 41: 751. 1932, Jour. Chem. Phys. 7: 339, 353, 356, 368. 1939.
[88] Polanyi, M., "Atomic Reactions." London. 1932.
[89] Rice, F. O., & Rice, K. K. "The Aliphatic Free Radicals." Johns Hopkins Press. Baltimore, 1935.

Propagation Processes

The simplest propagation reaction is the normal chain growth, which seems to take place fairly rapidly. Observations of Melville,[53] Schulz [16] and others [25] indicate that the average life time of a growing polystyrene chain of about 1000 monomers is between 10^{-3} and 10^{-2} seconds, if the radical mechanism is involved. On the other hand, Abkin,[25, 26] Bolland,[28] Medvedev,[25, 35] Norrish,[77] Raff [3] and Schulz [16] have observed that chains of butadiene, isoprene, acrylic- and vinyl-derivatives can grow over a very long period without being effectively terminated (polarized double bond mechanism).

The different chain transfer reactions are very interesting and important propagation processes. In the simplest case, chain transfer occurs according to the schedule

$$m_j^* + m_1 \longrightarrow m_j + m_1^*.$$

A growing chain with j monomers having a free valency at its end collides with a monomer, takes a hydrogen atom away from it, loses thereby its own radical character and transfers it to the monomer. Activation is maintained, but degree of polymerization is lost.

Uncatalyzed chain transfer has been postulated by Flory,[10] Schulz [16] and others [3] to explain the molecular distribution curves of certain vinyl-derivatives. Recently, Abere,[20] Breitenbach,[73] Goldfinger,[20] Maschin [73] and Naidus [20] have succeeded in investigating the catalyzed chain transfer during the polymerization of styrene in various solvents. They found that growing styrene chains attack CCl_4, add chlorine at their end and liberate a CCl_3 radical. This radical in turn starts a new chain, which continues to grow. This seems to explain in a fair way why certain solvents decrease the average molecular weight without affecting the rate of consumption of the monomer

Another propagation process is branching, which has been introduced in the discussion of styrene polymerization by Raff,[3] Schulz [15, 16] and Staudinger.[22] The first experimental evidence for the existence of branching was seen in characteristic discrepancies between degrees of polymerization as obtained from osmotic pressure and viscosity measurements, respectively. It seems, however, that this argument has lost its strength in view of more recent and more carefully collected data on the relationship between osmotic and viscosimetric data. Measurements by T. Alfrey,[19] A. Bartovics,[4] R. M. Fuoss,[6] P. J. Flory,[90] M Harris,[91] R. A. Kemp,[8] H. Peters,[8] H. A. Rutherford,[91] and A. Sookne [91]

[90] Lecture at the meeting of the New York Section of the American Chemical Society on February 5, 1943.

[91] **Rutherford, H. A., Sookne, A., Harris, M., & Mark, H.** Jour Res Natl. Bur. Standards **29**: 123. 1942.

indicate that a comparison of degrees of polymerization as obtained by the two different methods is a much more involved problem than it first seemed to be, and suggest other explanations for the above-mentioned discrepancies.

However, there exist other indications for the existence of chain branching in the case of polystyrene. R. Boyer[92] has recently used the electron microscope to compare polystyrene samples prepared at different temperatures and has found that they differ in aspect quite considerably. T. Alfrey[19] and A. Bartovics[4] have observed that the osmotic pressure of polystyrene solutions, which were made with samples prepared at 60°, 120° and 180° C., showed a distinctly different dependence upon concentration. Evaluating their findings with the theory recently proposed by P. J. Flory[93] and M. L. Huggins,[7] they conclude that samples prepared at different temperatures seem to have intrinsically different internal chain structures. This, of course, cannot be considered as being an experimental proof for the existence of branching, but it is an indication that there are fundamental differences between polystyrene samples which have been prepared at different temperatures.

Termination Processes

In the case of radical chain polymerization, it seems that termination occurs either if two growing chains collide with their activated ends (free valencies) or if one chain collides with the dissociation product of the catalyst. Abkin,[25] Bolland,[28] Medvedev[25, 35] and Melville[53] have studied the termination of growing chains of ethylene and butadiene derivatives by various substances and found in some cases (radical chain) that the chain ends are extremely sensitive to cessation, while in others (ionic chain) growing chains are surprisingly resistant to termination. However, all these observations are of a more qualitative character, and do not allow at present the drawing of very definite conclusions as to the exact mechanism of chain breaking.

[92] Discussion at The New York Academy of Sciences meeting on January 9, 1943.
[93] Flory, P. J. Ann. N. Y. Acad. Sci. 44: 419. 1943.

ELASTICITY AND FLOW IN HIGH POLYMERS

By Robert Simha
From Department of Chemistry, Howard University, Washington, D. C.

INTRODUCTION

A few summarizing remarks will be made on the molecular mechanisms responsible for the large elastic deformations obtained in rubber-like materials. The main part of this paper deals with rate phenomena which appear in amorphous materials in general and particularly in amorphous high polymers when one of the variables of state in the system, e.g., shearing stress, volume pressure or temperature, experiences a rapid change.

If an internal stress τ is created in a medium, a deformation γ is produced. Its magnitude and rate of change with time depend upon the magnitude of the stress:

$$\tau = \Phi\left(\gamma, \frac{\partial\gamma}{\partial t}, t_1 T\right),$$

where t is the time elapsed since the beginning of the experiment and T the absolute temperature. If τ is a function of γ (and T) only, then we speak of our material as an elastic solid, while dependence upon the rate of deformation (and T) only, characterizes the viscous liquid. These two cases therefore correspond to a steady state which is established at once, or at least at a rate much larger than could be measured in the course of an ordinary extension or shear experiment.

The elastic deformation of a crystal like NaCl involves the increase of distances between constituent particles against chemical forces. It leads to a value of the elastic modulus between 10^{10} and 10^{12} dyn cm^2 according to the nature of the forces involved. In this process only a change of distances occurs. However, each particle retains its original neighbors on the average. Viscous flow of a liquid, on the contrary, involves the displacement of particles relatively to each other and the occupation of new definite equilibrium positions. The energy gained in the course of this process is therefore dissipated into heat.

HIGH ELASTIC DEFORMATION

In the case of high polymers a different situation is encountered. Due to the size of the molecule, the rate at which a steady state is reached is

smaller, often considerably smaller. Furthermore, viscous flow is impeded in comparison with that exhibited by other amorphous substances while elastic deformation is favored. Besides the ordinary elastic deformation, characteristic of a perfect elastic solid, we find a second component of highly elastic deformation, exceeding the first one considerably. While the first type exhibits only slight temperature dependence, the opposite is true for high elasticity. According to the kinetic theory of rubber-like elasticity[1–5] this effect is connected with the uncoiling of molecules under the influence of the external stress and with the orientation of the chains in the direction of stress. The corresponding change in free energy is responsible for the appearance of a modulus of high elasticity of the order of magnitude of 10^6 dyn/cm^2 in a state of equilibrium. In the course of this process the density of the material remains approximately constant, indicating that the change in free energy is largely an entropy change. The heat motion opposes the orientation tendency and counteracts the external stress much as the external pressure is counteracted by the thermal motion of the molecules in a perfect gas.

Under certain simplifications these considerations can be evaluated by means of a statistical treatment to give numerical results for the equilibrium value of the high elastic component of deformation. The theory, as originally developed, contained the rather serious restriction of perfectly free rotation within one chain and the absence of any kind of interaction between chains. Recent work of Bresler and Frenkel[6] deals with the modification of the thermal motion and the shape of flexible chains by an internal potential hindering free rotation. Guth and James[7] have derived an expression for the equilibrium value of the high elastic modulus of a flexible network. In this manner it is possible to calculate stress-strain curves for rubber up to several hundred per cent extension.

The change in free energy upon stretching can be determined from thermal measurements. The equilibrium value of the tension τ is connected with the energy and entropy change on extension by means of well-known thermodynamic relationships[8]:

1 Wöhlisch, F. Jour. Soc. Biol. 87: 358. 1928.
2 Kuhn, W. Kolloid Zeit. 68: 2. 1934, 76: 258. 1936.
3 Guth, E., & Mark, H. Monatsh. Chem. 65: 93. 1934.
4 Frenkel, J. Acta Physicochim. USSR 9: 235. 1938.
5 Wall, F. T. Jour. Chem. Phys. 10: 132. 1942.
6 Bresler, S., & Frenkel, J. Acta Physicochim. USSR 11: 485. 1939
7 Guth, E., & James, H. M. Ind. Eng. Chem. 33: 624. 1941.
8 Wiegand, E., & Snyder, L. Trans. Inst. Rubber Ind 10: 234. 1934

(U internal energy, S entropy)

$$\tau = \left(\frac{\partial U}{\partial \gamma}\right)_T - T\left(\frac{\partial S}{\partial \gamma}\right)_T = \left(\frac{\partial U}{\partial \gamma}\right)_T + T\left(\frac{\partial \tau}{\partial T}\right)_\gamma \cdot \quad (1)$$

This gives the isothermal dependence of U and S on the extension γ. Wiegand and Snyder's[8] data on vulcanized rubber were evaluated by Treloar.[9] The result is shown in TABLE 1. Up to extensions of about 350 per cent there is only a small positive energy change on stretching and the larger part of the tension is due to a decrease in entropy. At higher elongations, however, crystallization begins, as reflected in the change of sign of $(\partial U/\partial \gamma)_T$, the rapid increase of all absolute values and the steepness of the stress-strain curve. This shows that high elastic deformation of moderately vulcanized, uncrystallized rubber is primarily an entropy effect and thus confirms at least in a qualitative fashion the basic assumption of the kinetic theory of rubber-like elasticity. In the case of plastics and fibrous materials the total apparent equilibrium deformation observed at room temperature is a much more complex superposition of energy and entropy effects, and crystal-like elasticity plays a much more important role even at moderate extensions. The analogy with a perfect gas breaks down and thermal motion is no longer the preponderant factor. One can conclude that rubber-like elasticity accompanies the existence of relatively free rotation within a chain as determined by the interactions along the main chain skeleton and by the nature and size of the side groups The investigations of Bresler and Frenkel[6] further indicate that the chain length is an important factor in determining the ability of the chain to assume shapes differing from the straight one at a given temperature, and to give rise to the above-mentioned entropy effects.

TABLE 1

INTERNAL ENERGY AND ENTROPY CHANGES ON STRETCHING

Extension γ %	Tension τ (g.wt.)	$(\partial U/\partial \gamma)_T$ (g.wt.)	$T(\partial S/\partial \gamma)_T$ (g.wt.)
158	70	+20	−50
288	104	+24	−80
376	121	+5	−116
462	113	−142	−255
548	131	−240	−371
632	156	−170	−326
718	250	+390	+140

[9] Treloar, L. R. G. Trans. Faraday Soc. 38: 293. 1942.

RATES OF DEFORMATION AND TIME EFFECTS

A distinction between rubber-like materials such as polyisoprene, neoprene or polyisobutylene and other high polymers is found not only in the absolute values for the apparent value of the elastic deformation but also in the rate of attainment of true equilibrium and of ultimate flow as given by the viscosity of the medium. The rate of these processes is determined by the velocity with which the molecules are able to react to the applied stress. This rearrangement is favored by the thermal motion, hindered by molecular forces, and is therefore strongly temperature dependent. The Maxwell-type[10] relaxation of stress at constant deformation, for instance, is caused by the stress-relieving transitions of constituent particles in the lattice into positions of lower free energy. The same mechanism leads, at constant tension, ultimately (or more correctly, after a time large compared with the relaxation time) to viscous flow. Alexandrov and Lazurkin[11] write the total elastic deformation in the following manner:

$$\gamma(t) = \gamma_{el} + \gamma_{hel}(\infty)\left[1 - \exp\left(-\frac{t}{\lambda}\right)\right] + \frac{t\tau}{\eta} \tag{2}$$

The last term has been added here to include viscous flow. η is the viscosity if γ denotes the shear. In the case of an elongation experiment, η is divided by $2(\mu + 1)$, where μ is the Poisson ratio and may be assumed to have a constant value if no measurable time effects are connected with the lateral contraction. Additional crystallization appearing in the course of the experiment is excluded for the moment. The first term reproduces the ordinary crystal elasticity, the second one the part that is characteristic of a high polymer. This relation can be understood in the following manner. Let us assume that the total deformation is composed of three parts: first, a term corresponding to the ordinary elastic extension and equal to τ/G, where G is the ordinary modulus of shear or of elasticity; second, a part reproducing viscous flow and equal to

$$\frac{1}{\eta}\int \tau\, dt;$$

last, a term accounting for that part of the reversible deformation which is established only gradually and, in agreement with equation (2), according to an exponential law. In other words, we assume as a first

[10] Maxwell, J. C. Phil. Mag. 35: 134. 18[illegible].
[11] Alexandrov, A. P., & Lazurkin, Yu. S. Acta Physicochim. USSR 12: 647 1940

approximation that the orientation process can be treated as a first-order reaction. The equation below satisfies this postulate:

$$\tau = G_1\gamma' + \eta_1 \frac{d\gamma'}{dt}$$

G_1 is the modulus of high elasticity in the case of high polymers, η_1 the viscosity connected with the establishment of final elastic equilibrium, such that $\eta_1/G_1 = \lambda$, defines the time necessary for this purpose. The total deformation can then be seen to obey the following integro-differential equation[12]:

$$G_1\gamma + \eta_1 \frac{d\gamma}{dt} = \left(1 + \frac{G_1}{G} + \frac{\eta_1}{\eta}\right)\tau + \frac{\eta_1}{G}\frac{d\tau}{dt} + \frac{G_1}{\eta}\int^t \tau\, dt; \qquad (2')$$

For constant stress τ it has equation (2) as solution if:

$$G\gamma_{el} = \tau; \qquad G_1\gamma_{hel}(\infty) = \tau.$$

Equation (2) can be given a somewhat different interpretation as will appear below. It may be seen that it will be valid only for small deformations, when the four parameters G, G_1, η, η_1 are really constants under isothermal conditions. Furthermore the inertia of the specimen has been neglected. This implies that the rate of change of the external stress is small compared with the frequency of the free vibrations of the sample. For a periodic stress $\tau e^{i\omega t}$ with angular frequency ω, equation (2′) gives a variable deformation with the same frequency and the following amplitude:

$$\gamma_{el} + \frac{\gamma_{hel}(\infty)}{1 + i\omega\lambda}\left(1 + \frac{\eta_1}{\eta} - \frac{iG_1}{\eta\omega}\right), \qquad (2a)$$

in which γ_{el} and $\gamma_{hel}(\infty)$ are defined as above. If flow is negligible, the term in the parenthesis reduces to 1. In any case, however the amplitude is complex. This, in turn, indicates a phase lag between the applied stress and the resulting strain analogous to the one observed between field strength and electric current in a dielectric which leads to dielectric losses. In the same manner mechanical loss in our medium is the result even if static internal friction is negligible. The amplitude of high elastic deformation is furthermore not a constant of the medium under consideration only, but depends also upon the magnitude of the stress frequency in a manner easily seen from equation (2a) and shown in FIGURES 6 and 7.

The meaning of $1/\lambda$ is that of a reaction rate if we consider a rate process in the general sense as a transformation requiring the over-

[12] Frenkel, J., & Obraztsov, J. Jour. Phys. Acad. Sci. USSR 2: 131. 1940.

coming of an energy barrier. The latter determines in this case mainly the parameter η_1. This viscosity coefficient must not be identified with the internal friction, η, which measures Newtonian flow in the sample. We can therefore write:

$$\lambda = A \exp \frac{\Delta H_{el}}{RT} \cdot \tag{3}$$

It is found[11] that for moderate strains, not exceeding 100 per cent in the case of partly vulcanized rubber for instance, this relation holds with a value of the heat of activation ΔH_{el} independent of the stress and also independent of the temperature in a range of about 25° C. for all materials considered. FIGURE 1 shows a logarithmic plot of λ for natural and

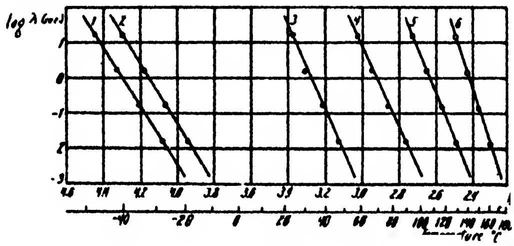


FIGURE 1. Variation of the relaxation time with temperature for different polymers, (1) rubber made of natural rubber with 3 per cent of sulfur, (2) chloroprene; (3) partly vulcanized ebonite; (4) methyl methacrylate with 30 per cent plasticiser; (5) the same with 10 per cent plasticiser added; (6) as before, without plasticiser.

synthetic rubber samples and methyl methacrylate samples with various plasticizer contents. Straight lines are obtained and it may be seen that the orientation time λ has the value of one second (an interval comparable with the time of loading) for rubbers at temperatures below −30° C., while the plastics do not reach such a value below +70° C. and the last unplasticized sample only at +140° C. The same facts are expressed in TABLE 2 of the heats and entropies for the process of rearrangement, computed by Eley[13] from Alexandrov and Lazurkin's[11] data. A variation in ΔH_{el} by a factor of 2 is found between the two limits of the table. FIGURE 2 shows the creep curve at various temperatures calculated from equation (2), neglecting the flow term. FIGURE 3 gives the influence of increasing temperature on the deformation for various times of loading computed from equations (2) and (3). FIGURES 4 and 5 contain the corresponding experimental data for natural rubber. FIGURE 6 shows the temperature dependence of the elastic deformation

[13] Eley, D. D. Trans. Faraday Soc. 38: 299. 1942.

TABLE 2

ENERGY AND ENTROPY OF ACTIVATION FOR ELASTIC ORIENTATION PROCESS

Polymer	ΔH_{el} kcal mol^{-1}	ΔS_{el} caldeg^{-1} mol^{-1}
Natural rubber—3% S	38	104
Chloroprene	38	99
Partly vulcanized ebonite	55	121
Polymethacrylate—30% plasticizer	52	87
Polymethacrylate—10% plasticizer	59	95
Polymethacrylate without plasticizer	75	122

at increasing frequencies according to equations (2a) and (3). These equations reproduce in a first approximation the experimental data for the polymers indicated in TABLE 2, as may be judged from FIGURES 4, 5 and 7. They show clearly that the stress-strain curve of a high polymer depends strongly upon temperature and experimental time or frequency in the case of a periodically varying stress. The higher the frequency at a given stress, the higher also the temperature at which a given strain can be produced. This is in accordance with the fact that even ordinary liquids with viscosities of the order of 10^3 10^4 poises assume solid-like properties; e.g., show brittleness,[14] or transmit shear waves, when exposed to mechanical disturbances of sufficiently high frequency.

The quantity characteristic for the time behavior at any temperature is the "orientation time," which in turn is determined by the Gibbs' free energy of activation for the process. The differences between a rubber and a plastic like polystyrene or an acrylic acid derivative seem to be constituted to a larger part by differences in energy of activation, while

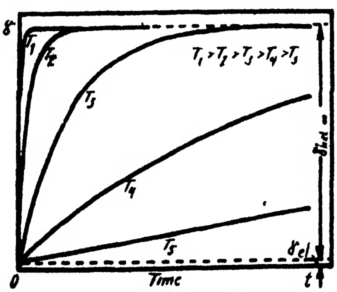


FIGURE 2. Deformation-time curves for different temperatures according to equation (2)

[14] Kornfeld, M., & Ryvkin, M. Jour. Phys USSR 2: 183 1940.

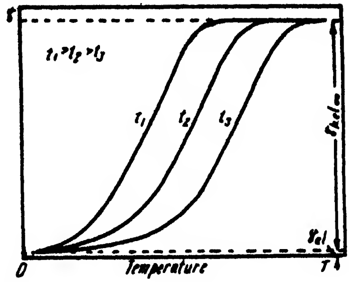


FIGURE 3 Deformation-temperature curves for various times t of loading, according to equations (2) and (3).

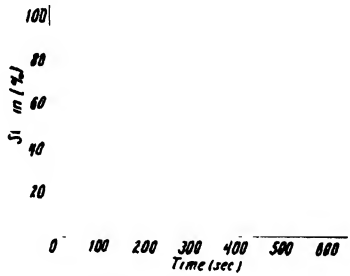


FIGURE 4 Creep curves for natural rubber (3 per cent S) at various temperatures

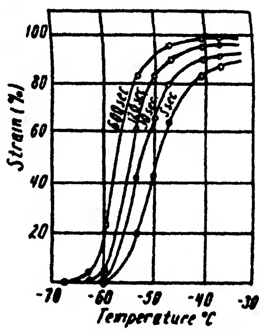


FIGURE 5. Deformation-temperature curves at various times of loading of natural rubber (3 per cent S).

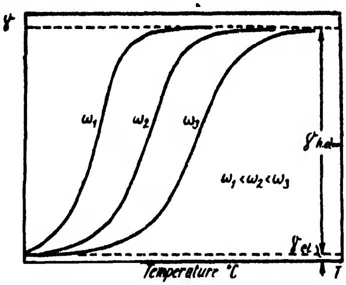


FIGURE 6. Deformation-temperature curves for various stress frequencies ω according to equations (2a) and (3).

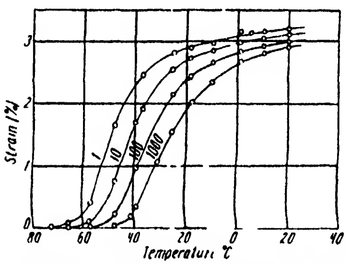


FIGURE 7. Deformation-temperature curves of natural rubber (3 per cent S) at varying frequencies.

entropy differences play a smaller role, as far as one can judge from TABLE 2. It must be kept in mind however, that these data refer to uncrystallized materials and to small deformations. ΔS will therefore contain mainly contributions of the change of shape of each chain and of the segment motions in the field of stress. This leads to another question, namely, how far is it justified to treat the whole viscous-elastic reaction as a simple kinetic process characterized by a single rate constant? Polar high polymers like phenolic resins[15] or polar vinyl derivatives[16] exhibit dielectric dispersion in a fashion which cannot be accounted for in terms of a single time constant. Instead, a whole distribution must be assumed, the character of which can actually be deduced from the nature of the dispersion curve.[16] It was furthermore shown by Fuoss and Kirkwood that a chain molecule gives rise to the existence of a multitude of dielectric relaxation times because of its internal flexibility which introduces the possibility of segment motions.

[15] Hartshorn, L., Megson, N. J. L., & Rushton, E. Proc. Phys. Soc. 52: 817. 1940.
[16] See for instance Fuoss, R. M., & Kirkwood, J. G. Jour. Am. Chem. Soc. 63: 385. 1941. Jour. Chem. Phys. 9: 329. 1941.

The nature of this work shows that a similar situation will exist in respect to mechanical behavior.[17] An extension of Maxwell's[10] original relaxation theory to a relaxation spectrum was first proposed by Thomson[18] and Wiechert.[19] Kuhn[20] has attributed high elasticity to the existence of different molecular mechanisms, with different elastic moduli and widely differing time constants coming into play when stress is applied. The nature of the creep curves obtained with inorganic glasses[21] and high polymers[22] points in the same direction.

The broadness of the spectrum in the case of a high polymer is essentially a consequence of the particular effects characteristic of long-chain molecules possessing internal flexibility of shape. Three different mechanisms can be roughly distinguished as responsible for the mechanical time effects.[23] First, the thermal diffusion of the chain segments in the field of shear. Although not identical with their movements in an electric field this leads to relaxation times of the same order of magnitude as those responsible for dielectric dispersion; that is, to values below one second. Next we must consider the change in average shape caused by the application of the stress. The free energy of activation for the relaxation of stress is larger here The relaxation time increases with the size of the molecule and with its internal stiffness. For an isolated chain with free rotation the time is proportional to the degree of polymerization and inversely proportional to the rate of diffusion of one chain end relatively to the other.[4] Hindered rotation increases the value[6] as is to be expected. The van der Waals forces between chains will have the same effect. The alignment and displacement of whole chains relatively to each other requires the overcoming of van der Waals forces between whole chains, hence the values of the time constants are large. Flow will become preponderant after a period, t, large compared with these times, which are of the order of magnitude of hours or days in a fiber. Obviously the above viscosity, η_1, is to be correlated to the first and second mechanism since, for instance, in rubbers at room temperature, it is smaller than η by orders of magnitude.

According to these considerations the time constant λ in equation (2), if it is valid at all, must represent an average over the whole distribution. We can obtain a very rough estimate of the limits in the following fashion. A relation of the type of equation (2) has been derived by various

---

[17] This is confirmed by data of **Ponamarev, L. T.** Jour. Tech. Phys. USSR **10**: 588. 1940.
[18] **Thomson, J. J.** "Applications of Dynamics to Physics and Chemistry." London, 1888.
[19] **Wiechert, E.** Ann. Physik. Chem. **50**: 335. 1893.
[20] **Kuhn, W.** Zeit. physikal Chem **B42**: 1. 1939.
[21] **Taylor, N. W.** Jour. Appl. Phys. **12**: 753. 1941.
[22] For instance: **Leaderman, H.** Textile Research **11**: 171. 1941.
[23] **Simha, R.** Jour Phys Chem **47**: 348. 1943

authors[24, 25] by assuming the existence of a discrete number of relaxation mechanisms, each following Maxwell laws. This corresponds to a parallel arrangement of spring-dash-pot pairs coupled in series, or in other words to a superposition of stresses rather than strains which led to equation (2′). For the particular case of two such relaxation times the following relation is found:

$$\gamma(t) = \frac{\tau}{E_1 + E_2}\left\{1 + \frac{E_1E_2(k_1 - k_2)^2}{(E_1k_2 + E_2k_1)^2}\left[1 - \exp\left(-\frac{E_1k_2 + E_2k_1}{E_1 + E_2}t\right)\right] + \frac{k_1k_2}{E_1k_2 + E_2k_1}\cdot t\right\}.$$

E_1 and E_2 are elastic or shear moduli, according to the case under consideration, and k_1 and k_2 represent the rates of relaxation for the two mechanisms. This is identical with equation (2). The inhomogeneity of the relaxation process gives rise to a series of exponential terms instead of a single one if more than two independent relaxation times are introduced. Of course similar results can be obtained by an extension of equation (2′). We are using this approach here because it has already been extended[26] in a way similar to the work of Fuoss and Kirkwood[16] on dielectrics.

It may be seen that the equilibrium value of the high elastic deformation will be larger, the larger the disparities in the values for the two relaxation rates at a given temperature. For our estimate we assume that viscous flow is negligibly small. This implies that one of the rates, say k_2, is vanishingly small. It follows that the ratio between high elastic and ordinary elastic deformation equals the inverse ratio of the elastic moduli. We may assign a value of 10^{11} dyn/cm² to E_1, and 2×10^6 dyn/cm² to E_2. At a temperature of −40° C., the λ-constant for rubber has the approximate value of one second. Taking this average value and our previous assumptions, we find $k_1 = 5 \times 10^4$ sec^{-1} and $k_2 = 0$. At −25° C., k_1 would have increased to 5×10^{11} sec^{-1}, if the heats of activation remained constant over this wide range of temperatures. At room temperature, the methacrylate sample with 30 per cent plasticizer has a value k_1 of 5 sec. with $k_2 = 0$, and therefore exhibits no rubber-like behavior. The procedure used can furnish only a very rough estimate, especially at higher temperatures where viscous flow will be measurable. Nevertheless, it shows that several relaxation mechanisms of widely differing magnitudes are operating which cannot

[24] **Bennewitz, K., & Rötger, H.** Physikal. Zeit. 40: 416. 1939
[25] **Holzmüller, W., & Jenckel, E.** Zeit. physikal. Chem. A186: 359 1940
[26] **Simha, R.** Jour. Appl. Phys. 13: 201. 1942.

be represented by an average value. The creep curves obtained lead to similar conclusions and they can be used for a determination of the actual distribution of relaxation times and of the Boltzmann after-effect function.[27,26,23]

Oppression of long-chain effects, by increasing the van der Waals forces, for instance, in the transition from a rubber to a plastic or fiber,[28] will not only shift the spectrum to smaller frequencies but also narrow its width and reduce the extent of high elastic deformation. Cross linking will increase the largest relaxation times and therefore retard the onset of flow though not affecting appreciably the elastic properties, as long as the number of cross bonds is small compared with the number of intra-molecular links. TABLE 2 shows that ultimately cross bonding (ebonite) increases the mean energy of activation for elastic deformation. This must be ascribed primarily to the part of the energy necessary for segment motions. The higher values obtained for the methacrylates may be due to the introduction of polar side groups. Eley[13] has attempted to separate the tabulated energies ΔH_{el} into two parts, one to be attributed to the separation of the chain segments in the process of orientation which leads to high elastic deformation. The corresponding value of ΔH is determined from the activation energy of viscous flow in raw rubber which is about 10 kcal per mole, and from the energy necessary for rotation of one C–C bond which has the order of magnitude of 3 kcal. The remainder of the activation energy of 38 kcal is ascribed to the rotation of C–C bonds necessary for the uncoiling of the chains. This is analogous to the previously discussed separation of the relaxation spectrum into two parts, if viscous flow is negligible. In view of the uncertainty of the energy values for internal rotation and of the fact that the activation energies are not additive in averaging over the relaxation distribution, these conclusions appear questionable.

RELATION TO OTHER PHENOMENA

The ordinary elastic modulus and the shear modulus are practically independent of temperature. For the other coefficients the kinetic theory predicts an increase proportional to a first or higher power of the temperature[6] according to whether the thermal energy or the rotation potential around valence bonds is larger. This dependence is overshadowed by the influence of temperature on the time constant. The attainment of the final value of the highly elastic strain is increasingly retarded as the temperature is lowered at constant stress. Finally

[27] Boltzmann, L. Pogg. Ann. Erg. 8: 624. 1876.
[28] Mark, H. Ind Eng. Chem. 34: 1343 1942.

a situation is reached in which this equilibrium value is not established at all for all practical purposes. It may be seen, for instance, from the extension curves previously shown, that for natural, moderately vulcanized rubber this happens at a temperature of about $-70°$ C. Above this transition temperature the total equilibrium modulus is determined by the equation:

$$1/G_{\text{total}} = 1/G + 1/G_1 \approx 1/G_1.$$

Below it a nonequilibrium with a higher value of the free energy is frozen in with a modulus given by:

$$1/G_{\text{total}} \approx 1/G.$$

These facts bear a great resemblance to the volume pressure and volume–temperature behavior observed in low molecular glasses as well as in high polymers. Tammann and Jellinghaus[29] found that selenium glass and colophonium frozen originally under atmospheric pressure assumed a smaller volume under a pressure increase only upon heating the specimen to a certain temperature T_f. Similarly, the volume change accompanying a reduction of pressure could be observed only under the same condition. An analogous situation exists in respect to thermal expansion.[30] Extensive measurements on rubbers, plastics and low molecular materials have been carried out by several authors. Jenckel[31] has measured the rate of volume change for the substances investigated by Tammann and his collaborators. The position of the transition point varies of course with the rate of temperature change. By extending the duration of the experiment from less than an hour, as done by Tammann and Kohlhaas, to five hours, the transition temperature was reduced from 30° C. to 23° C.

The volume change under isobaric or isothermal conditions is exactly analogous to the change of strain discussed previously. An instantaneous part results from the change of the amplitude of thermal vibrations, just as in an ordinary crystal. That is, the average volume occupied by a molecule is changed. This component is the analogue to ordinary elastic extension. Second, a reorientation takes place, in the course of which the mutual arrangement of the molecules as a whole and of their parts, if we are dealing with high polymers, is altered. In other words, the distribution function of equilibrium positions in the system, which describes the short range order existing in liquids or amorphous

---

[29] **Tammann, G., & Jellinghaus, E.** Ann. d. Phys. **2**: 264. 1929.

[30] **Tammann, G., & Kohlhaas, A.** Zeit. f. anorg. u. allg. Chem. **182**: 49. 1929. **Jenckel, E., & Überreiter, K.** Zeit. physikal. Chem. **A182**: 361. 1938. **Überreiter, K.** Zeit. physikal. Chem. **B45**: 361. 1940. **B46**: 157. 1940; **Wiley, F.** Ind. Eng. Chem. **34**: 1052. 1942.

[31] **Jenckel, E.** Zeit. Elektrochem. **43**: 796. 1937.

solids, passes to a new equilibrium value over an energy barrier which, together with the temperature, determines the time necessary for the accomplishment of this process.

These reactions to changes in the different variables of state are related to each other, because of the similarity in the molecular mechanisms. TABLE 3, presented in part by Tuckett,[32] shows that the transition

TABLE 3

COMPARISON OF TRANSITION TEMPERATURES FOR ELASTIC MODULUS AND THERMAL EXPANSION

Polymer	T_F	Approx. Range over which High Elasticity Develops
Unvulcanized rubber	−66° C.	−50–−70° C.
Polyisobutene	−65° C.	−50–−70° C.
Polystyrene	+80° C.	+70–+90° C
Polymethacrylate	+60–+124° C.	+70–+110° C.

temperature T_f for the thermal expansion actually lies within the range T_e over which the transition from crystal-like to high elasticity develops. The introduction of cross links or of plasticizer[30] changes T_f in a similar fashion as T_e. No pressure-volume data on high polymers have come to the author's attention, but one can expect an analogous situation to prevail, as indicated by Tammann's work.

The relation between the relaxation of shearing stress or of shearing deformation and of pressure or relative volume change can be obtained approximately from the well-known equation connecting tensor of stress and tensor of deformation rate in a viscous liquid. In a compressible fluid we have for the mean normal stress p acting on a surface element:

$$p = \mu_1 \operatorname{div} \vec{v} - p_s,$$

where p_s denotes the hydrostatic pressure and $\vec{v}$ the velocity vector. μ_1 is the volume viscosity which measures the stress set up if the specimen is uniformly compressed at a given rate.[33] For incompressible systems (that is, in the ordinary applications of hydrodynamics, where a single constant, the shearing viscosity, characterizes the medium) the divergence of $\vec{v}$ vanishes. The mean stress p is then simply equal to the hydrostatic pressure with the opposite sign. Introducing the compressibility κ_1, and remembering the physical significance of the diver-

[32] Tuckett, R. F. Trans. Faraday Soc. 38: 310. 1942.
[33] Lamb, H. "Hydrodynamics" Cambridge. 1932

gence as a relative volume change, we obtain the following expression relating pressure and volume variation:

$$\delta p = \frac{1}{\kappa_1}\frac{\delta V}{V} + \mu_1 \frac{d}{dt}\frac{\delta V}{V}. \tag{4}$$

Superposition of a volume change connected with "crystal-like" compressibility then leads to an equation analogous to equation (2′) and to the same conclusions as for shear. At high temperatures or low compression frequencies a large value of the total equilibrium compressibility obtains and a discontinuity results below a certain transition temperature.[12] The characteristic time is determined by the product of volume viscosity μ_1 and compressibility κ_1 which will be practically equal to the equilibrium compressibility above the transition point. It is therefore possible to determine the volume viscosity by following the time dependence of volume upon pressure change. In view of the values obtained for the compressibilities of high polymers,[34] the volume viscosity μ_1 will be larger in general than the shearing viscosity η_1. It would be very interesting to have accurate data on the absorption and velocity of propagation of sound waves in a frequency range,[12,35] since such measurements would give a value for the volume viscosity.[36] The second order transition in the compressibility would of course appear in these measurements.

Similar considerations apply to the isobaric volume change. The previous role of viscosity is played in this case by an inverse heat conductivity which, however, does not bear any direct connection with the ordinary constant. A systematic investigation of all three rate phenomena on identical samples along the lines of Tammann and Jenckel is very desirable. An indication that the actual volume changes may not follow the simple exponential laws resulting from the assumption of two different molecular mechanisms is given in Jenckel's[31] work on selenium. It is found that a logarithmic plot of volume gives a straight line versus the square root of time instead of the time itself. It may be noted that similar statements have appeared in the literature in regard to creep curves. On general thermodynamic grounds one will furthermore expect a correlation between the second-order transition as observed

[34] See for rubber hydrocarbons the summary of Wood, L. A. Proc. Rubber Technology Conference London 1938. P 933

[35] Mandelstam, L., & Leontowič, M. Jour Expt. and Theor. Phys. 7: 438. 1937, Isaakovich, M. Compt. rend. Acad Sci. USSR 23: 783 1939

[36] See in this connection the interesting work of Ferry, J. D. Jour. Am Chem. Soc. 64: 1323. 1330. 1942.

for instance in various rubbers by Bekkedahl and collaborators[37] on the basis of heat capacity data, and the transitions discussed previously.[38]

We may conclude these remarks by pointing out one particular limitation of the results presented here. Equation (2) can be valid only for systems in which no first-order phase change occurs in the course of the experiment; that is, crystallization processes must not appear on increase of deformation. It would not be difficult to arrive at a modification of the creep curve by making plausible assumptions about the connection between actual deformation and degree of crystallinity as derived from X-ray data. However, there does not seem to be sufficient quantitative information available to induce such an approach. It should furthermore be possible to develop a theory that is free from such additional assumptions and gives both the relation between deformation and amount of crystallization and the creep curve as the resultant between rate of crystallization and rate of relaxation. In such a theory the results presented here and related ones would appear as limiting cases for vanishing rate of crystallization.

[37] **Bekkedahl, N., & Matheson, H.** Jour. Res. Natl. Bur. Standards 13: 411. 1934, **Bekkedahl, N., & Scott, R. B.** Jour. Res. Natl. Bur. Standards 29: 87. 1942.

[38] Compare the work on hydrocarbons by **Van Hook, A., & Silver, L.** Jour. Chem Phys. 10: 686. 1942.

THE RIGIDITIES OF SOLUTIONS OF POLYMERS

By John D. Ferry

From the Department of Physical Chemistry, Harvard Medical School, Boston, Mass.

INTRODUCTION

Solutions of a linear polymer of high molecular weight in a solvent with which it is miscible in all proportions grade uniformly from viscous liquids through gelatinous or rubber-like consistencies to plastic solids. At moderately high concentrations in the liquid range, no rigidity can be observed in a static experiment. However, such solutions will support brief impulsive stresses or rapidly oscillating stresses, revealing small but measurable rigidities. At higher concentrations, in the neighborhood of equal parts of polymer and solvent, static experiments yield rigidities approaching those of rubber-like solids, and dynamic experiments give somewhat higher values still. Finally, as the concentration of polymer approaches 100 per cent, the static modulus of rigidity rises to a value characteristic of either a rubber-like solid (10^6 dyn cm^{-2}) or a hard solid (10^{10}–10^{12} dyn cm^{-2}), depending on the temperature and the magnitude of intermolecular forces between polymer chains. The dynamic modulus of rigidity attains a value characteristic of a hard solid in any case. In general, these transitions of viscous to rubbery and rubbery to solid properties are shifted to lower concentrations by decreasing the temperature, or by increasing the frequency in dynamic experiments.

The number of polymeric solutions whose rigidities have been measured is small, and no single system of polymer and solvent has yet been studied over the entire concentration range. After some remarks on theory and experimental methods, the existing data will be reviewed. Although incomplete, they permit some generalizations to be made.

THEORY

The modulus of rigidity, G, is defined[1] as

$$G = \mathfrak{T}/\gamma \tag{1}$$

where $\mathfrak{T}$ is the shearing stress producing an angular deformation, or strain, γ.

[1]Timoshenko, S. "Theory of Elasticity." McGraw-Hill Book Co New York, N. Y. 1934. Coker, E. G., & Filon, L. N. G. "Photoelasticity." Cambridge University Press 1931.

The appearance of rigidity in polymeric solutions is accompanied by several complicating phenomena:

(*a*) Relaxation of stress at constant strain;
(*b*) Viscous flow at constant stress, in series with elastic deformation;
(*c*) Delayed elastic recovery after cessation of viscous flow;
(*d*) Variation of rigidity with frequency.

These phenomena can be related by various theoretical models, two of which have been described in the preceding paper by Simha.[2] The simplest model is the Maxwell concept[3] of viscous deformation in series with elastic deformation. From this the relaxation of stress at constant strain (*a*) is given as

$$\mathfrak{T} = \gamma G e^{-t/\tau}$$

where τ is a mechanical relaxation time, and the viscosity (*b*) is $\eta = G\tau$. A single Maxwell element cannot describe delayed elastic recovery (*c*). It does, however, describe the frequency variation of rigidity (*d*), as follows.

The velocity of propagation of transverse vibrations in a medium whose relaxation time is long compared with the period of the vibrations is

$$V = (G/\rho)^{1/2} \tag{2}$$

where ρ is the density. However, when the period of vibrations is the same order of magnitude as the relaxation time, the rigidity increases with increasing frequency in this range. For a plane wave whose amplitude can be expressed as $\gamma = \gamma_0 e^{i(\omega t - 2\pi x/\lambda) - x/x_0}$, and for a single relaxation time, the dispersion of rigidity is given[4,5] by

$$\frac{G(\omega)}{G^\circ} = \frac{V^2(\omega)}{(V^\circ)^2} = \frac{2\omega^2\tau^2}{\omega^2\tau^2 + \omega\tau\sqrt{1 + \omega^2\tau^2}} \tag{3}$$

and the critical damping distance, x_0, is related to the relaxation time by the formula

$$\tau = \pi x_0/\omega\lambda - \lambda/4\pi\omega x_0, \tag{4}$$

where ω is the circular frequency and λ the wave length.

The utility of the Maxwell concept is much increased by representing the relaxation of stress by a series of exponential terms, each with its own contribution to rigidity and characteristic relaxation time.[6,7] Such

---

[2] Simha, R. Ann. N. Y. Acad Sci. 44: 297. 1943.
[3] Maxwell, J. C. Phil. Trans. Roy. Soc. 157: 52. 1867, Phil Mag. (4) 35: 133. 1868.
[4] Derjaguin, B. Beitr. angew. Geophysik 4: 452. 1934
[5] Ferry, J. D. Jour. Amer. Chem. Soc. 64: 1323. 1942.
[6] Kuhn, W. Z. physik Chem. B42: 1. 1939.
[7] Bennewitz, K., & Rötger, H. Physik. Z. 40: 416. 1939

a model can describe all four of the phenomena mentioned above, but its relation to molecular behavior is obscure. The utility of other models for the mechanical behavior of polymeric solutions will be discussed below.

EXPERIMENTAL METHODS

Measurement of the rigidity of a solution of a polymer involves two difficulties. First, there is usually a large viscous displacement in series with the elastic displacement, and the two must be evaluated separately. Second, the measurement requires a finite time to carry out, and any contributions to the rigidity associated with relaxation times smaller than that interval will be lost.

When the relaxation times are of the order of minutes or longer, a static experiment will suffice, and correction for viscous displacement can be made by extrapolating back from a state of steady flow or by measuring the recovery after removal of load. Materials too weak to support their own weight can be sheared between coaxial cylinders, the outer serving as a container.[8, 9, 10]

When relaxation times are less than a second, however, it is necessary to use either impulsive or alternating stresses with periods shorter than the relaxation times. For materials strong enough to support their own weight, vibrations can be produced mechanically or electrically and the relation between stress and displacement determined. This has been performed recently by Alexandrov and Lazurkin[11] and by Sack.[12] The phase angle between stress and displacement can be measured, thus determining both elastic and viscous components. For weak materials, this is more difficult. In the apparatus of Kendall,[13] the sample is subjected to a shearing impulse of very short duration (10^{-5} sec.), and the resulting displacement and recovery followed by rapid photographic recording. In the apparatus of Philippoff,[14] the sample is contained in a cup and an electrically driven needle dips into it, vibrating in its own axis. From the driving force and the observed displacement, the viscosity is calculated as a function of frequency; from its dispersion, a relaxation time is determined, and the rigidity is taken as the ratio of the static viscosity to the relaxation time. For weak materials that are transparent and show strain double refraction, the propagation of transverse vibrations can be studied with an apparatus described by the

[8] Schwedoff, T. Jour. de physique (2)8: 341. 1889.
[9] Hatschek, E., & Jane, R. S. Kolloid-Z. 39: 300. 1926.
[10] Ferry, J. D., & Parks, G. S. Physics 6: 356. 1935.
[11] Alexandrov, A. P., & Lazurkin, Y. S. Acta Physicochim. USSR 12: 647. 1940.
[12] Sack, H. S., reported at the Society of Rheology, New York, October 31, 1942
[13] Kendall, J. M. Rheology Bulletin 12: 26. 1941.
[14] Philippoff, W. Physik. Z. 35: 883. 1934.

author.[15] Measurement of the velocity of propagation yields the rigidity (equation 3) independent of any viscous displacement. It should be remarked that the methods of impulsive and alternating stress usually give an *adiabatic* modulus.[16]

By combining static and dynamic methods, especially over a range of frequencies, the contributions to the rigidity associated with various relaxation times can be segregated.

Another means of determining a modulus of rigidity is from the dependence of viscosity on shearing stress in certain solutions. For polystyrene in xylene,[17] the apparent viscosity η' follows the equation

$$\eta' = \eta_0/(1 + \mathfrak{T}/G_a) \tag{5}$$

The quantity G_a, which has the dimensions of a modulus of rigidity, is proportional to the concentration and can be interpreted as a contribution to the modulus associated with long relaxation times.[18]

RESULTS

The systems which have thus far been studied include solutions of polybutene, polystyrene, polyvinyl chloride, and polymethyl methacrylate.

Polybutene

An isolated determination by Kendall,[13] using a shearing impulse of 10^{-5} sec. duration, gives the modulus of rigidity of a dilute solution of polybutene in pseudocumene as 10^4 dyn cm^{-2}. The concentration is not given, but the molecular weight was 80,000 and the viscosity 16 poises, which probably corresponds to a concentration of several per cent.

The propagation of transverse vibrations in concentrated solutions of polybutene in heptane has been studied by the author.[19] The average molecular weight of the material was 385,000. There was no dispersion in the frequency range of 320 to 1250 cycles. The modulus of rigidity increased with increasing temperature, and was in fact proportional to the absolute temperature.[20] At 25° C., it was 2.6×10^4 dyn cm^{-2} at a con-

[15] **Ferry, J. D.** Rev. Sci. Inst. **12**: 79. 1941.
[16] **Guth, E.** Private communication.
[17] **Ferry, J. D.** Jour. Amer. Chem. Soc. **64**: 1330. 1942.
[18] For emulsions of starch in a mixture of paraffin and carbon tetrachloride (**Hatschek, E., & Jane, R. S.** Kolloid-Z. **40**: 53. 1926), **Philippoff, W.**, (Kolloid-Z. **71**: 1. 1935) demonstrated a somewhat different relation, $\eta' = \eta_0/(1 + \mathfrak{T}^2/G_a^2)$. Here, also, G_a is proportional to the concentration and interpreted as a modulus of rigidity.
[19] **Ferry, J. D.** Unpublished work.
[20] The variation with temperature suggests that this rigidity is due to strain orientation entropy (**Kuhn, W.** Kolloid-Z. **76**: 258. 1936; **87**: 3. 1939.) However, this is probably fortuitous. The variation with concentration is evidence that this rigidity is due to intermolecular transfer of stress, as explained in the next section. The temperature coefficient of such a rigidity, like that of the viscosity of polymer solutions (**Alfrey, T., Bartovics, A., & Mark, H.** Jour. Amer. Chem. Soc. **64**: 1557. 1942) may be expected to be either positive or negative, and sensitive to the relative intermolecular forces of solvent and polymer.

centration of 13.4 per cent polymer, and 1.3×10^5 at 23.6 per cent. On the basis of these two values, the rigidity appears to vary with slightly less than the third power of the concentration.

Polystyrene

Solutions of polystyrene in xylene, in the concentration range of 15–52 per cent, have been investigated by the author.[5, 17] The average molecular weight of the material was 120,000. The modulus of rigidity was calculated from the dependence of viscosity upon shearing stress, from the elastic recovery following shear under constant load, and from the propagation of transverse vibrations.

The modulus G_a calculated from the anomalous viscosity (equation 5) was independent of temperature within experimental error. It was proportional to the weight fraction g up to over 30 per cent polymer, following the equation $G_a = 1.1 \times 10^4\, g$ dyn cm^{-2}. At higher concentrations it rose more rapidly; at 52.3 per cent, it was 1.2×10^4.

The modulus calculated from elastic recovery, determined only for the 52.3 per cent solution, was 1.0×10^4 dyn cm^{-2}, in good agreement with the above figure. The recovery process was not described by an exponential relation, but could be characterized by an average relaxation

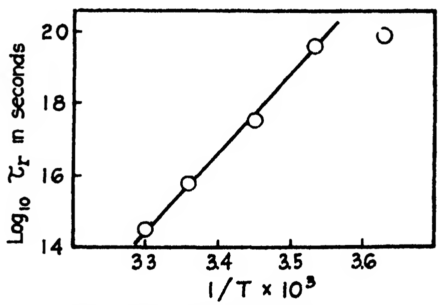


FIGURE 1. Mean recovery time of 52.3 per cent solution of polystyrene in xylene, plotted logarithmically against the reciprocal absolute temperature

time (at which $(e - 1)/e$ of the recovery was accomplished), which varied from 28 sec. at 30.5° to 98 sec. at 2.7°. The temperature variation of this relaxation time (FIGURE 1) followed the equation $\tau = Ae^{Q/T}$, with $Q = 9300$ cal., a value close to that characterizing the temperature dependence of the viscosity of this solution (10,700 cal.).

The modulus calculated from the propagation of transverse vibrations

underwent dispersion at the lower concentrations, and the high-frequency values G° (equation 3) were determined by extrapolation.

This dynamic modulus was considerably higher than the static modulus described above, and ranged (at 25°) from 0.75×10^4 dyn cm^{-2} at a concentration of 15 per cent polymer to 38×10^4 at 52.3 per cent (FIGURE 2). Its dependence upon temperature followed the equation $G = Ae^{Q/RT}$, and the heat effect Q was 1500 cal., independent of concentration up to 40 per cent (FIGURE 3).

The relaxation time was derived from the dispersion data at the lower concentrations (equation 3) and from measurements of damping at the higher concentrations (equation 4). It was found to be 4×10^{-4} sec., independent of concentration and temperature, within experimental error.

On the basis of a set of Maxwell elements,[6, 7] these data can be interpreted in terms of two rigidity mechanisms, setting $G^\circ = G_1 + G_2$ and $\eta = G_1\tau_1 + G_2\tau_2$. If G_1 be identified with the modulus calculated from anomalous viscosity and from elastic recovery, and τ_2 be identified with the relaxation time derived from studies of transverse vibrations, then (using values of the viscosity measured for these same solutions[17]) the remaining quantities τ_1 and G_2 are readily calculated (TABLE 1). The

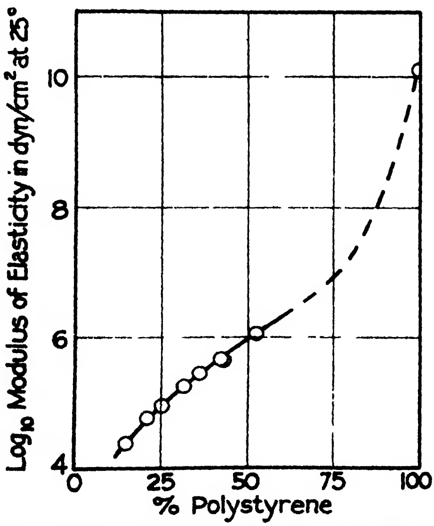


FIGURE 2. Modulus of elasticity of the system polystyrene-xylene.

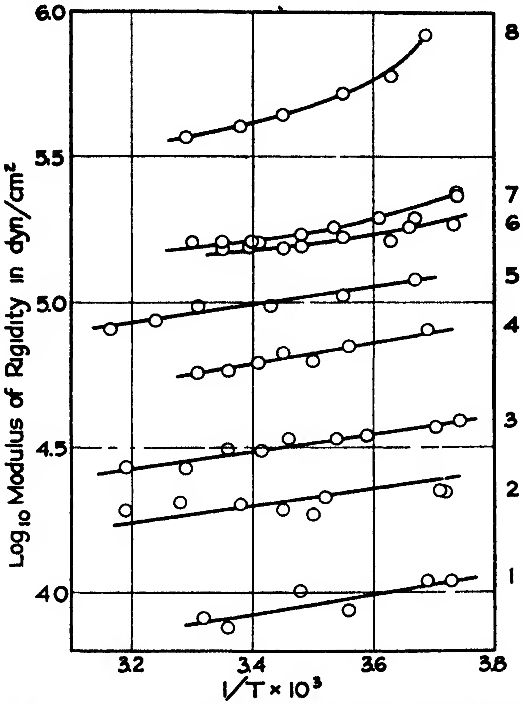


FIGURE 3 Rigidities of polystyrene-xylene solutions, plotted logarithmically against the reciprocal absolute temperature 1, 15 3 per cent, 2, 20 6 and 21 3 per cent, 3, 25 7 per cent, 4, 31 8 per cent, 5, 36 0 per cent, 6, 42 5 per cent, 7, 42 2 per cent, 8, 52 3 per cent

modulus G_2 varies with the third power of the concentration, up to over 30 per cent (FIGURE 4), and the time τ_1 (which must be considered the average of a rather wide distribution of relaxation times) is also strongly concentration dependent.

TABLE 1

ANALYSIS OF RIGIDITY MECHANISMS IN POLYSTYRENE-XYLENE, AT 25° C

Concentration Polystyrene, per cent	$G^\circ \times 10^{-4}$	$G_1 \times 10^{-4}$	$G_2 \times 10^{-4}$	η	$\tau_2 \times 10^4$	$\tau_1 \times 10^2$
15	0.75	0.16	0.59	13.1	4.0	0.73
20	1.66	0.22	1.44	52.5	4.0	2 12
25	2.82	0.28	2.54	209	4 0	7.1
30	4.80	0.33	4.47	660	4.0	19.5
35	9.34	0.39	8.95	2,000	4 0	50
52.3	38	1.2	36.8	116,000	4.0	950

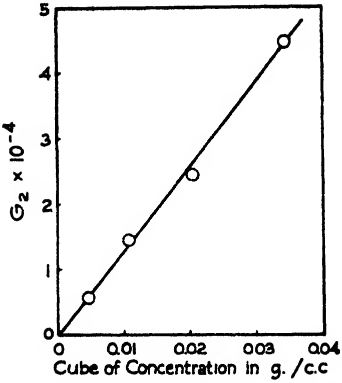


FIGURE 4. Modulus of rigidity of the second mechanism (G_2) in polystyrene-xylene solutions plotted against the third power of the concentration.

On the basis of these mechanisms, it is of interest to depict the variation of rigidity over a wide frequency range. Using the values in TABLE 1, the logarithm of the rigidity (at 25°) is plotted against the logarithm of the frequency in FIGURE 5. At the lower concentrations, the frequency range in which mechanism 1 operates without mechanism 2 is quite small. At 52.3 per cent polymer, however, it is a plateau which extends over several decades. It is only when this plateau reaches to frequencies of 1 sec.$^{-1}$ and less, of course, that obvious rubbery properties make their appearance.

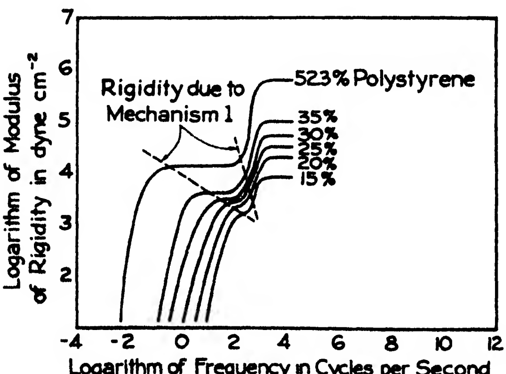


FIGURE 5. Modulus of rigidity, at 25°, of the system polystyrene-xylene, based on the constants of TABLE 1.

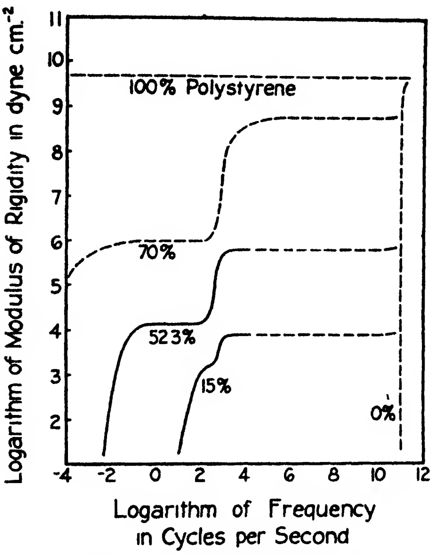


FIGURE 6. Modulus of rigidity, at 25°, of the system polystyrene-xylene (schematic)

A schematic extension of FIGURE 5 can be made by drawing in curves for the pure polymer[21] and the solvent (FIGURE 6). The latter is based on the observation of Raman and Venkateswaran[22] of rigid behavior of liquid glycerol at a frequency of 10.[10] The critical frequency for xylene may be expected to be higher.[23] Also, a purely hypothetical curve is drawn for the rigidity of a plastic with 70 per cent polystyrene. The inflection of this curve represents the well-known "brittle point" of long-chain polymers, which is a transition[11] from a modulus characteristic of a rubbery solid (10^6–10^7 dyn cm^{-2}) to one characteristic of a hard solid (10^{10}–10^{12} dyn cm^{-2}). The interpretation of the two rigidity mechanisms in terms of molecular processes will be discussed below.

Polyvinyl Chloride

Solutions of polyvinyl chloride have been studied in the range of 40–100 per cent polymer. Davies, Miller and Busse[24] have reported measurements of the modulus of elasticity of polyvinyl chloride plasticized with tricresyl phosphate, using an experimental time of 5 seconds

Calculated from Young's modulus as given by the manufacturer, assuming Poisson's ratio to be 0.5

[22] **Raman, C. V., & Venkateswaran, C. S.** Nature **143**: 198. 1939.

[23] **Goodeve, C. F.** Trans. Faraday Soc. **35**. 342. 1939.

[24] **Davies, J. M., Miller, W. A., & Busse, W. F.** Jour. Amer. Chem. Soc. **63**. 361. 1941.

Their figures, expressed as the logarithm of the modulus of rigidity (in dyn cm^{-2}), are given in TABLE 2. Here Poisson's ratio is assumed to be 0.3 for the three highest concentrations, and elsewhere 0.5.

TABLE 2*

MODULUS OF RIGIDITY OF POLYVINYL CHLORIDE–TRICRESYL PHOSPHATE

Per cent Polymer	Log G		
	−10°	32°	70°
40	7.40	6.63	6.63
50	8.59	7.09	6.89
60	9.07	7.40	7.07
70	9.76	8.59	7.29
80		9.32	7.51
90		9.47	8.35
100		9.61	9.35

* Data of Davies, Miller, and Busse.

On a graph of the type of FIGURE 6, these points, corresponding to a frequency of 0.2 $sec.^{-1}$, lie in the upper left corner. At the lower concentrations, they doubtless represent plateaus at a rubbery level, like that hypothetically drawn for 70 per cent polystyrene. At the highest concentrations, they represent the modulus of a hard solid, the critical frequency of the transition from rubbery to hard having shifted to much lower values. The influence of increasing temperature in shifting the transition to higher concentrations is apparent.

Similar data of Busse reported by Fuoss,[25] for polyvinyl chloride plasticized with diphenyl (FIGURE 7), demonstrate the transition from a rubbery solid ($G = 10^{7.0}$ dyn cm^{-2}) to a hard solid ($G = 10^{9.3}$). As before, increasing the temperature shifts the transition to higher concentrations (lower concentrations of plasticizer).

Polymethyl Methacrylate

The mechanical behavior of plasticized polymethyl methacrylate was studied by Alexandrov and Lazurkin,[11] who investigated the effect of concentration, temperature and frequency. No calculations of rigidity were made, but the plot of relative strain (FIGURE 8) shows the familiar transition from rubbery to solid behavior. At 40°, the critical frequency lies somewhat below one cycle for a concentration of 70 per cent polymer, and far below this for the higher concentrations. From the temperature dependence of the critical frequency, the energies associated with the

[25] Fuoss, R. M. Jour. Amer. Chem. Soc. 63: 378. 1941.

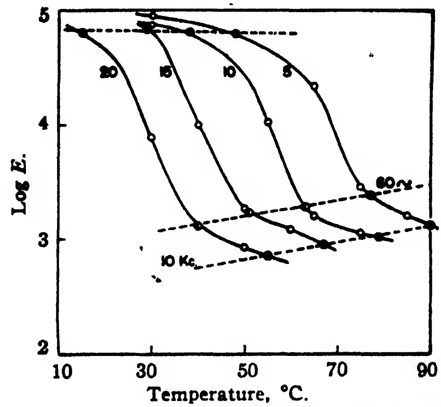


FIGURE 7 Modulus of elasticity of the system polyvinyl chloride-diphenyl, plotted logarithmically against the temperature (Fuoss, R. M., Jour. Amer. Chem. Soc., 63. 378. 1941). The figures opposite the curves refer to percentages of diphenyl.

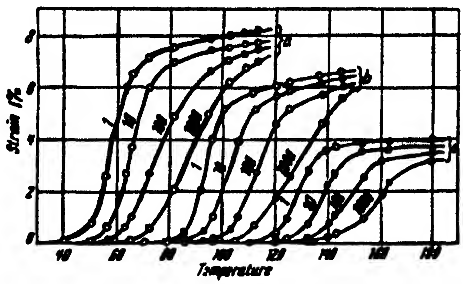


FIGURE 8. Strain plotted against temperature for plasticized polymethyl methacrylate (Alexandrov, A. P., & Lasurkin, Y. S., Acta Physicochimica USSR, 12: 647. 1940); (a) 70 per cent polymer; (b) 90 per cent polymer; (c) 100 per cent polymer. The figures opposite the curves refer to frequencies in cycles per second.

relaxation process have been calculated[2] for each concentration and are quite large, e.g., 52 kcal. at a concentration of 70 per cent. The moduli of rigidity at 180° (in the rubbery range) are in the ratio 0.45 : 0.58 : 1 for concentrations of 70 per cent, 90 per cent and 100 per cent polymer.

DISCUSSION

The concept of two mechanisms for support of shear as used for polystyrene-xylene systems may be applied also to the other solutions except at very high concentrations. It is interesting to attempt to relate these mechanisms to molecular processes.

The first is associated with a broad distribution of relaxation times. For experimental periods longer than these times, the solution is a liquid. The average relaxation time varies markedly with concentration —from a small fraction of a second in dilute solutions to minutes or hours in the range of equal parts of polymer and solvent. The modulus of rigidity first increases linearly with concentration (in polystyrene-xylene), then more rapidly. This mechanism is supposed to be the same as that usually held responsible[26] for rubber-like elasticity– namely, the orientation entropy of long-chain molecules. Relaxation of stress here may take place by movement of whole polymer chains or long segments of them. The temperature dependence of elastic recovery is the same as that of the macroscopic viscosity; and probably viscous flow, elastic recovery, and stress relaxation of this first rigidity mechanism all involve the same kind of molecular motion.

The second mechanism appears to be associated with a sharp relaxation time in dilute solutions. In more concentrated solutions, the form of the dispersion function is not known. The relaxation time is independent of concentration up to 50 per cent in polystyrene-xylene. It is independent of temperature at lower concentrations. The modulus of rigidity first increases with the third power of concentration, then more rapidly. This mechanism is supposed to be the bending of carbon–carbon bonds in molecules immobilized between "points of entanglement," where the long chains cross. The number of these points should increase very rapidly at high concentrations, until a closely entangled structure is attained in the pure polymer. The relaxation process here may involve very short segments of chains, moving relatively freely between neighboring molecules. Such a segment, confined between two other chains but able to oscillate between them, would essentially move in a medium of low viscosity until it collided with one of its neighbors and formed a point of entanglement.

As the concentration of polymer approaches 100 per cent, a third mechanism probably contributes to the rigidity—the intermolecular attractions between neighboring polymer chains. These attractions no doubt play a greater role in polymers with polar groups, like polymethyl methacrylate and cellulose esters, than in hydrocarbons like polybutene and polystyrene. Whether the modulus of rigidity of an undiluted polymer is determined primarily by attractive forces between chains or by bending of bonds in mechanically enmeshed, closely tangled molecules, is not clear.

[26] **Guth, E., & Mark, H.** Monatsh. für Chemie 65: 93. 1934, **Mark, H.** Jour. Appl. Phys. 12: 41. 1941; **Kuhn, W.** Kolloid-Z. 76: 258. 1936, 87: 3. 1939.

Representation of Mechanical Behavior by Models

Combinations of springs and dashpots have often been used to represent the mechanical behavior of systems possessing both rigidity and viscosity. The utility of such a model depends on its ability to describe all of the four phenomena listed on page 314, and on the possibility of identifying its elements with actual molecular processes.

A pair of Maxwell elements, as employed in interpreting the polystyrene-xylene system following Kuhn and Bennewitz and Rotger, corresponds to two parallel sets of a spring and dashpot in series (FIGURE 9, A). This model represents viscous flow and frequency dependence with the values assigned to the constants shown in the figure and in TABLE 1, but it fails to describe the elastic recovery correctly. According to the model, since G_2 relaxes quickly and the recovery consists in G_1 acting against the viscosity $G_2\tau_2$, the recoil time should be $G_2\tau_2/G_1$, which is 0.012 sec. for 52.3 per cent polystyrene at 25°. The average time observed is 38 sec. Furthermore, this arrangement of mechanical elements is not reasonable in terms of the molecular mechanisms postulated. If one source of deformation is the uncoiling of randomly kinked molecules and the other is bending of carbon–carbon bonds along the chains, these two deformations certainly occur in series and not in parallel. A reasonable model must show the two rigidity elements, G_1 and G_2, in series

The second model drawn (FIGURE 9, B) fulfils this requirement. It corresponds to Simha's[2] equation (2), which is a modification of the equation of Alexandrov and Lazurkin,[11] the series viscosity having been added to account for viscous flow. This model, applied to the polystyrene data, represents not only viscous flow and frequency dependence but also the elastic recovery. If η^* is set equal to η, the macroscopic viscosity, the average recoil time for 52.3 per cent polystyrene at 25° is η/G_1 = 9.7 sec., which is comparable with the observed time of 38 sec. Also, the fact that the temperature variations of the recoil time and of the macroscopic viscosity yield the same activation energy follows from this relationship.

However, the model of FIGURE 9, B does not explain dispersion of rigidity with a relaxation time (τ_2) of the order of 10^{-4} sec. in the polystyrene solutions. To account for this, a new mechanical element can be introduced, forming the model shown in FIGURE 9, C. Here we have in series with the previous elements a dashpot whose viscous resistance is much smaller than η and is numerically equal to $G_2\tau_2$. The displacement of the dashpot is limited by a mechanism, drawn as a chain in the diagram, so that the viscosity of the whole system has the low value $G_2\tau_2$

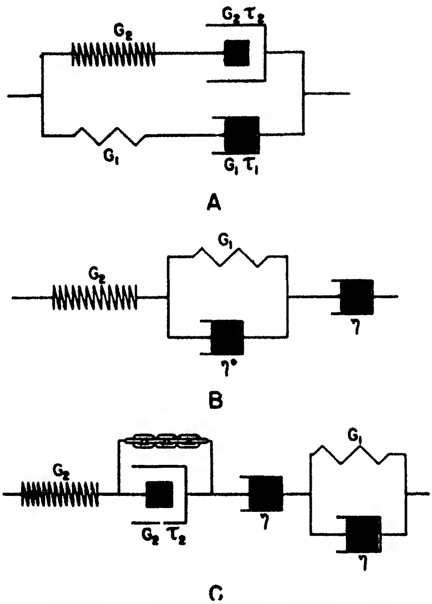


FIGURE 9. Models for the mechanical behavior of polymeric solutions: (A), two Maxwell elements in parallel, (B), model for the equation of Simha (equation 2, reference 2), (C), model proposed to describe the system polystyrene-xylene.

only for small deformations. For large deformations, the viscosity is η. Thus the dispersion of rigidity at small deformations with a relaxation time τ_2 is accounted for, as well as viscous flow under constant stress, and elastic recoil, which is the same as in the previous model.

In the model of FIGURE 9, C, each element can be directly interpreted in terms of molecular behavior. The rigidity G_2 is the resistance to bending of carbon–carbon bonds in segments of molecular chains between points of entanglement where chains cross. The viscosity $G_2\tau_2$ is the resistance to motion of short-chain segments which are confined between other polymer molecules but with enough spacing to allow some rattling back and forth with comparatively little interference up to the point of collision with one of the neighboring chains. The point of collision (point of entanglement) is represented by the chain being stretched taut across $G_2\tau_2$, and thereafter with continued extension the entire viscosity η is encountered. This represents the motion of entire molecules, or long segments of them, being dragged along by the points of entanglement. Finally, the rigidity G_1 is that due to strain orientation entropy,

and it is reasonable to show this in parallel with the macroscopic viscosity if the orientation involves motion of entire molecules or long segments of them.

Thus the model of FIGURE 9, C is quite satisfactory for the existing data, but may require revision after further experiments—on the dependence of viscosity on amplitude, for example. As more complete data are accumulated for a variety of polymeric solutions, there will be many different ways of describing the mechanical properties by models or by analytical expressions. The most useful descriptions will be those which can be identified with molecular processes.

SUMMARY

1.—The theory of the mechanical properties of solutions of high polymers is discussed in terms of the simplest mechanical model, the Maxwell element.

2.—Several methods of determining the rigidities of solutions of polymers are described.

3.—The results of measurements of the rigidities of solutions of polybutene, polystyrene, polyvinyl chloride, and polymethyl methacrylate are reviewed.

4.—The mechanical properties of these systems are discussed in terms of molecular behavior, with special reference to solutions of polystyrene in xylene.

5.—A mechanical model is proposed which describes all the properties of the polystyrene-xylene system, and consists of elements which can be identified with molecular processes.

INTERMOLECULAR FORCES AND CHAIN CONFIGURATION IN LINEAR POLYMERS—THE EFFECT OF N-METHYLATION ON THE X-RAY STRUCTURES AND PROPERTIES OF LINEAR POLYAMIDES

By

W. O. Baker and C. S. Fuller

From the Bell Telephone Laboratories, New York, N. Y.

INTRODUCTION

If linear polymer molecules are assumed to have a relatively high, fixed, average molecular weight, the properties of their solids depend primarily on two factors: the kind and extent of the molecular order or arrangement present, and the intermolecular forces. The linear polyamides have been chosen as models to investigate these relations, and the general results apply to all chain polymers.

The molecular order may approach that of crystalline perfection as it does in ramie and certain synthetic polyesters,[1] or it may reach maximum disorder as in unstretched, unfrozen natural rubber. In general, real polymeric compounds, unless they are completely disordered molecularly, consist of a dispersion of two or more kinds or degrees of order,[2] and these are associated with various chain configurations.[3] With X-ray patterns as a criterion, it is convenient to distinguish regions which show: (1) crystalline order in which the repeating units are arranged according to a lattice in three dimensions; (2) mesomorphic order in which the units are arranged parallel to the chain direction with corresponding groups opposite but are otherwise random or in which the units are simply arranged parallel to the chain direction; and (3) no molecular order. In this case the repeating unit loses its significance and the order becomes atomic. These regions may be termed amorphous.

Certain subclasses of mesomorphic order exist for essentially linear polymers. One possible description of these follows:

Type I.—Lateral disorder of chains of uniform composition (homochains). That is, the chains are composed of a single repeating unit. Their lateral disorder means that they are randomly rotated about their long axes, in the polymer solid. In addition, of course, they are presumably kinked, but we are considering more local order, less affected

[1] Carothers, W. H. "Collected Papers." Interscience Publishers. New York. 1940. Fuller, C. S. Chem. Rev. 26: 143. 1940.

[2] Gerngross, O., Herrmann, K., & Abitz, W. Zeit. physikal Chem. B10: 371. 1930.

by gross orientation, than that of the meandering paths of long chains. Examples of this class include polyamides,[5] cellulose triesters,[3] polyvinylidene chloride, polyethylene, etc. in the quenched state.

TYPE II.—Longitudinal disorder in chains of nonuniform composition (heterochains). Here the chains may include repeating units of various lengths, so that equivalent chain sections, such as polar linkages, do not come together in adjacent, parallel chains. Evidently this order may also have superimposed on it type I above[4] or type III below. Examples are copolyamides, copolyesters, copolysulfides and indeed most copolymers. However, many vinyl and diene copolymers have such large side groups that the effects of the type II disorder is obscured by extensive type III steric disorder. Also, chain inversion,[4] in which adjacent chain segments do not have the repeating units running in the same direction along the chains, might cause special longitudinal disorder.

TYPE III.—Steric disorder from side groups in either homo- or heterochains. Chains made irregular in packing by side groups include most vinyl polymers, such as polyvinyl chloride, polystyrene, etc. Here the side groups do not occur regularly or frequently enough (as contrasted to polyisobutylene, which is quite highly ordered when stretched) so that the combined main-chain and side-group packing is orderly. Also, vinyl copolymers as polyvinyl chloride acetate, synthetic rubbers such as Bunas, cellulose partial esters and mixed esters, natural proteins and many other chain polymers show this disorder.

It is assumed always that the given order refers to the properties of the solid polymer under definite conditions, i.e., stretched rubbers or quenched polymers[3,4,5] may have temporary degrees of order.

The preceding three classes of order are not proposed to classify uniquely the organization in a given polymer solid. Rather, they are supposed to show the varieties of chain configurations and packing which must be considered in investigating a definite chemical composition.

The N-methyl substituted polyamides discussed in this report possess chiefly type III disorder, which leads also to type I disorder. The methyl groups occur in the dipole layers of the solids and disorganize the forces in these regions. Other disorganization of these dipole layers has been found in the copolyamides.[4] In both cases hydrogen bonding is greatly reduced, in analogy to the substitution of cellulose to give esters, ethers, etc. Further, significant differences in configuration along the chains, apparently involving a twisting or bending, seem to

[3] ...r, W. O., Fuller, C. S., & Pape, N. R. Jour. Am. Chem. Soc. 64: 776. 1942.
[4] Baker, W. O., & Fuller, C. S. Jour. Am. Chem. Soc. 64: 2399. 1942.
[5] Fuller, C. S., Baker, W. O., & Pape, N. R. Jour. Am Chem. Soc. 62: 3275. 1940.

occur as a result of this force weakening and becoming disordered.[4] Since such twisting is suspected in the mechanism of high elasticity, direct evidence for it bears on the properties of rubbers. The highly substituted N-methyl polyamides indeed show rubberiness. X-ray diffraction has shown that various chain configurations in these compounds do exist as functions of external stress and degree of methylation (force weakening). Other physical properties, such as elastic modulus and moisture sorption, have also been measured and related to the molecular structure.

EXPERIMENTAL DATA

Materials

Polydecamethylene sebacamides in which the degree of N-methylation was varied from 0 to 55 mol per cent of the amide groups were employed. They were prepared from purified intermediates and were all of sufficiently high average molecular weight to allow ready drawing of fibers for X-ray examination. In the range up to 50 per cent methylation, usually the monosecondary diamine was employed to supply the methyl components. Above 50 per cent a random mixture of disecondary and monosecondary diamines was employed. No effect of the distribution of methyls among the primary amide groups was evident from these experiments.

Samples were oriented from quenched[5] or partially quenched polymers, and were then annealed in an inert atmosphere for 2 hours at 20° to 25° ± 0.5° C. below their melting points. This caused no discernible disorientation, but in all cases the expected improved crystallization. The samples were 23 to 25 mils thick when exposed, unless they were further stretched. In most cases they were annealed in a clamp to prevent shrinkage, and where noted they were further elongated after annealing, and thus exposed in a small clamp which was rigidly fastened on the X-ray collimator. The jaws of this clamp were movable so that the samples could be slowly and uniformly stretched.

X-Ray Diagrams

X-ray photographs of the fibers and unoriented sections were obtained as described previously.[4] A CA-6 G. E. tube supplied Cu K-radiation through beryllium windows. The K_α was filtered with Ni foil to remove the K_β. Salt dusted on the mono-filaments gave direct distance calibration. The specimen-to-plate distance was commonly 6 cm. and critical features were repeatedly checked. Fibers were frequently irradiated in

positions 90° apart about the fiber axis, to establish uniaxial orientation. The doubly-oriented (biaxial) samples were produced by drawing and rolling of filaments.

Elastic Modulus

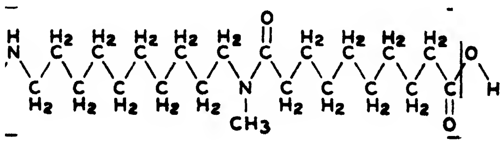
Methods previously discussed[4] were used to obtain Young's modulus, E, from the penetration of a quartz spherical section into a plane panel of the polymer. However, some of the 10–10 (the numbers refer to the carbon atoms in the acid and diamine chains) N–CH_3 polyamides were so soft that even light loads caused a slight permanent deformation; hence, some of the values quoted reflect a plastic as well as elastic behavior. The polymer panels were dried before test, and average values from triplicate readings are reported.

Moisture Sorption

The water sorption was determined on flat sections 3 × 2 × 0.051 cm. Equilibrium was checked by repeated weighings after exposure at 25.0° C. to saturated ammonium sulfate (81% RH) and immersion in water (100% RH). The sorption is reported as weight per cent of the dried (phosphorous pentoxide) sample. Humidified samples were dried and re-equilibrated, and only a very small hysteresis was found.

RESULTS AND DISCUSSION

The effects of polar coordination and disorder on mechanical properties of solid polymers already have been examined.[4] FIGURE 2 illustrates the sevenfold change in stiffness which was introduced in the polydecamethylene sebacamide series by substitution of various (molar) proportions of methyl groups for hydrogen atoms in the diamine. The structure for a typical composition is shown in FIGURE 1. This wide range in properties in this case apparently is chiefly the result of a sharp



AVERAGE UNIT OF
50% N-METHYLATED 10-10 POLYAMIDE

FIGURE 1. Structural formula of an average unit of polydecamethylene sebacamide with 50 mol per cent of N-methylation.

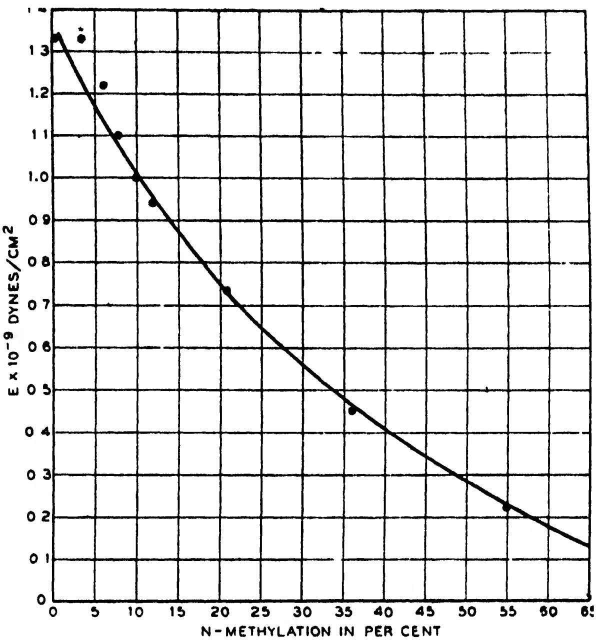


FIGURE 2. Dependence of Young's modulus, E, on the molar proportions of N-methylation for polydecamethylene subacamides.

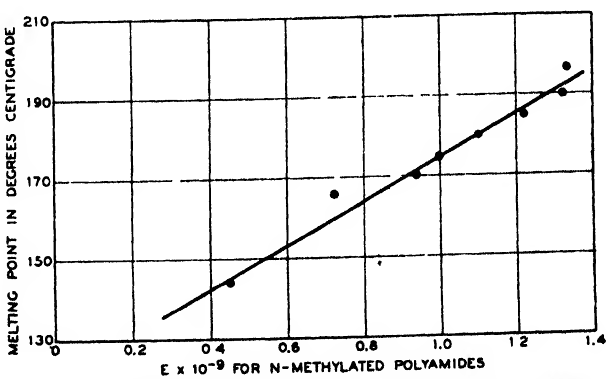


FIGURE 3. Relation of the elastic modulus to the melting point of N-methylated polydecamethylene sebacamides.

reduction in intermolecular forces. Methylation removes hydrogen bonding and reduces dipole attraction sterically, much as do larger acyl radicals in cellulose esters.[3] Thus, while the unmethylated polyamide is hard, high melting and somewhat brittle, the higher ranges of methylation are soft and rubbery. If reduced molecular interaction causes this transition in mechanical properties, the melting points should follow the decline of elastic modulus in the series. This is shown in FIGURE 3, where the relation is approximately linear.

The elastic modulus, whose values appear in TABLE 1, approximates an exponential function of composition, of the form $E = Ae^{-Bc}$, where A and B are constants, and c is the mole fraction of N-methylated groups.

TABLE 1

10–10 Polyamide *N-Methylation %*[a]	$E \times 10^{-9}$ (*dynes/cm.*^{2})	Water Sorption	
		81% RH	*100% RH*
0	1 33	1.6	2.0
3.5	1.30	1.6	2.2
6	1.22	1.6	2.1
8	1.10	1.6	2.1
10	1.00	1.6	2.1
12	0.94	1.7	2 3
21	0.73	1.7	
36	0.45	2.0	2.6
55	0.22	2.4	3.1

[a] Van Slyke.

Small polar molecules can be sorbed on polar linkages which are not strongly associated. Thus, the moisture sorption of the polyamides would be expected to rise with increasing methylation, as shown in FIGURE 4 and TABLE 1. This is another indication of reduced intermolecular forces in the solids. It is interesting that the moisture sorption increases by about 50 per cent of its original value when about half of the –NH– groups are methylated. Unlike the copolyamides,[4] water take-up is here uncomplicated by changes in polar group concentration. In all cases, however, it can serve as a measure of the organization in the dipole layers. The hydrogen bonds of perfectly ordered, associated –CO.NH– groups are apparently seldom split by water at 25° C.; for example, in films, the self-bonded linkages are about 640 cal./mole more stable than those bonded to water.[6] While moisture may be sorbed in the dipole fields of self-bonded linkages, it will be expected chiefly to seek amido groups uncoordinated because of chain disorder or of N-meth-

[6] Alexander, A. E. Proc. Roy. Soc. A179: 470. 1942.

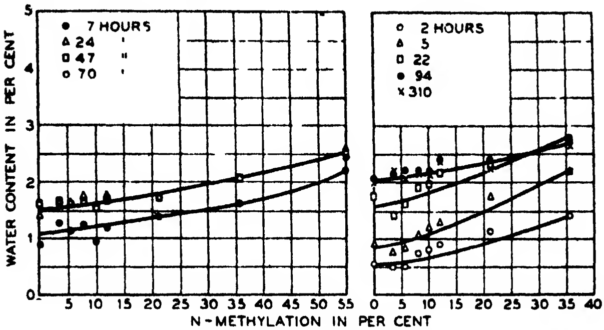


FIGURE 4. The water sorption at 81 and 100 per cent relative humidity, respectively, of N-methyl substituted polydecamethylene sebacamides.

ylation. Similar effects occur in partially substituted cellulose and proteins. There remains, of course, the further factor that among comparable hydrophilic materials, softer substances always absorb more water than harder ones. This must be considered in interpreting sorption results.

Although unsubstituted, hydrocarbon-chain, polyamides are soluble only in strong hydrogen-bonding solvents such as cresol, the reduction in interchain forces discussed above should increase the solubility in a given series. Thus, while 40 per cent ethanol–60 per cent chloroform (by volume) only penetrates the usual polyamides, such a mixture swells the lower ranges of methylation, and dissolves the higher. FIGURE 5 illustrates this solvation, since it indicates the amount of extraction (of low species) from comparable samples in contact with this solvent at 25° C.

The chemical measurements therefore agree with the mechanical studies in implying intermolecular forces weakened by substitution. The elastic modulus values show, for example, that the polymer chains (in segments) may be displaced from potential minima more easily the higher the substitution, or lower the effective polar group association.[4] The question occurs of whether such displacements can cause significant variation in inter- or intra-chain configurations. This is one of the primary problems in the theories of rubberiness. Since, as noted above, much rubberiness appears in the disordered polyamides, their structures yield information from X-ray diffraction on the behavior of chain molecules under stress.

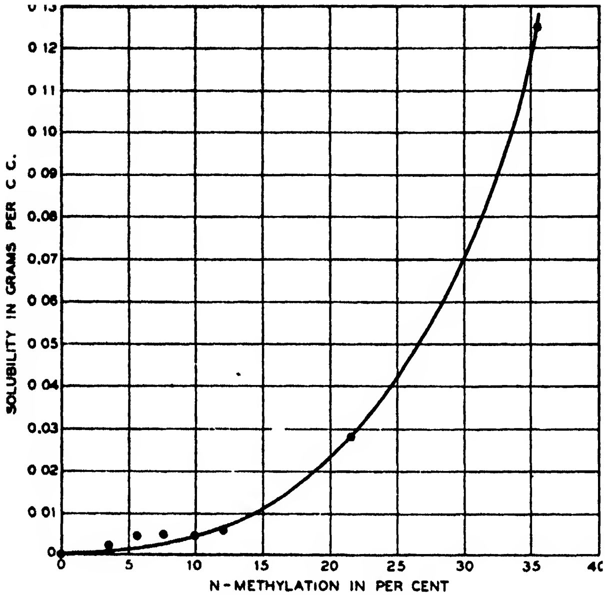


FIGURE 5. The relative solubilities in ethanol-chloroform mixture of polydisperse samples of N-methylated polydecamethylene sebacamides of comparable molecular weights.

Inter-Chain Spacings

The equatorial or near-equatorial fiber features yielded in all cases of annealed fibers the characteristic[4,5] polyamide spacings of 3.76 and 4.40 Å. These are taken to represent lateral chain separations in the stable, crystalline chain configurations of polyamides whose dipole layers are oblique to the chain axes. When these same polyamides are quenched, there is metastable dipole association into planes that are perpendicular to the chain axes. Here the single equatorial spacing is about 4.18 Å, suggestive of hydrocarbon chain packing. This same spacing is found in certain polyamides (as from base units containing odd-numbered chains, such as 9–9), whose chains assemble perpendicular to the dipole layers even in the most crystallized state. These variations are found throughout the N-methyl polyamides, but the methyl side groups do not appear to alter directly the lateral separation of the chains. They do diffuse some of the equatorial features at the higher degrees of methylation. Since main chain axes do not approach much closer than 4.7 Å, there is evidently room for a methyl group of bond

radius 1.47/2 Å and kinetic theory radius 2 Å. Thus, while the amido hydrogen projects from the N for a bond distance of 1.0 and a probable kinetic theory distance of 1.1 Å beyond that, for a total of 2.1 Å, the methyl group probably projects 3.5 Å from the N. To its lack of a bonding hydrogen there is thus added, for the N-methylated linkage, a steric hindrance which may lead to chain repulsion. Hence, it is interesting that in the present series the prime effect of a lateral substituent is to change *not* the side spacings but the long spacings of the chains. These long spacings have previously been found for many of the simple polyamides[4] to be less than the calculated values. For instance, polydecamethylene sebacamide generally exhibits an identity period, I, of 25.6 Å; the expected value for an extended zigzag chain is 27.5 Å. Further, for the following methylated polyamides, and various others not discussed, I values can be reduced by 26.5 per cent of the calculated values.

Relation of Mechanical Treatment to Fine Structure

PLATE 1,A is for the lowest per cent of N-methylation, 3.5. This pattern is typical of polyamide fibers. The quenched polymer was cold-drawn, fixed in a clamp, and annealed for 2 hours at 170° C. in an inert atmosphere. PLATE 1,B is the strikingly altered diagram obtained from a portion of the same fiber annealed under like conditions without a clamp. The normal shrinkage[5] of the fiber has been attended by a change in structure. The identity period spacing has decreased from 25.8 Å for the clamped specimen to about 20 2 Å for the retracted form. The outer equatorial spots of 1,A have split in 1,B, and the resolved layer-line features have coalesced into intense meridian reflections. The phenomenon is the same in all of the substituted polyamides and related substances in which it is observed. TABLE 2 compares typical observed features of the patterns for the 3.5 per cent methylated fibers.

Since the final crystalline forms of PLATE 1,A (extended) and 1,B (retracted) resulted from different mechanical treatments of the original oriented but highly disordered (quenched) fiber, the latter must contain molecular configurations capable of crystallizing in the solid state in different forms. PLATE 1,C is a typical quenched fiber pattern,[5] in this case for 9.9 mole per cent N-methylation. The long chains are approximately parallel to the fiber axis and are randomly rotated about their long axes; their dipole planes are nearly normal to the axis and in these planes the polar linkages are strongly associated. The single equatorial spot of PLATE 1,C corresponds to about 4.18 Å; the principal meridian

TABLE 2

Specimen—3.5% methylation	Reflection	Inter-Planar Spacing, A
Fiber clamped during annealing	A_1	4.4
	A_2	3.8
	II_0	25.8
	II_1	25.9
Fiber free during annealing	A_1	4.5
	I_0	20.2
	II_0	20.3
	III_0	19.8
	IV_0	19.8
	III_1	20.1

feature, II_0, to 24.8 Å. This distance is thus between the extended and retracted values. The usual behavior on annealing polyamides with relatively high polar group concentration and thus high interaction, such as polyhexamethylene adipamide and polyhexamethylene sebacamide is like the clamped behavior of the less strongly associated polymers. Thus, PLATE 1,C would be transformed to PLATE 1,A. It now appears that this transformation of the substituted polyamides *unclamped* can be effected by annealing at low temperatures. The freedom of the molecules in the solid is then insufficient for formation of the retracted state. PLATE 1,D is from annealing at 150° C. for two hours a 5.7 per cent methylated sample, *unclamped.* While the outer equatorial spots show a definite elongation tending toward the retracted forms, the identity period from II_1 layer lines is 23.6 and from the II_0 meridian spots, 25.8. Both are well above the retracted value of 20.2, but mixed forms are also indicated. However, a form of molecular clamping has been seemingly introduced by the lower temperature crystallization.

It will be shown below that the retracted spacings can always be transformed into extended by some sort of stretching. In the experiment above, can the apparently stable "intermediate" form represented by PLATE 1,D be similarly influenced? PLATE 1,E is the sample of 1,D stretched 130 per cent at room temperature. The II_0 features now represent an I value of 26.4 Å; the equatorial spacings are unchanged and the arcing of the outer spots has been narrowed. It now appears that the process of apparent chain elongation (to give a value only 4 per cent rather than 26.5 per cent shorter than the calculated) has drawn the formerly tilted (PLATE 1,D) dipole layers into a position approximately normal to the chain axes. This is not a low energy arrangement,[7] and therefore the fiber has to be photographed under stress.

[7] Baker, W. O., & Yager, W. A. Jour Am. Chem. Soc. 64: 2171. 1942.

But this experiment emphasizes further[4] that extended forms having the calculated planar chain spacing value are not stable, for these compounds.

Although the oriented state has so far been considered, the occurrence of the retraction phenomenon is not limited by this condition. PLATE 1,F is a typical diagram from an unoriented quenched section. The ring corresponding to the identity period has a d value of 24.3 Å. When such a sample is annealed (for instance, a 9.9 per cent N-methylated specimen), the pattern of 1,G results, and here the spacing along the chain from ring d values is 21.3 Å. The same figures were obtained from the Debye-Scherrer rings of samples allowed to crystallize very slowly from the melt as illustrated by PLATE 1,H. Such spacings correspond to a 22.5 per cent shortening of the calculated period. Thus, retracted chain forms[7a] occur spontaneously in the formation of a random mass of crystallites. These forms presumably represent the lowest free energy and resemble the conditions assumed for unstretched rubber.

Effect of Stretching

Since the crystallization of disordered polyamide chains may be diverted into the extended form by clamping or low temperature annealing, the question occurs of whether the retracted form can be converted by stress alone. Certain effects of stretching polyester fibers have been noted[8] which resemble those found here, and probably have similar causes. FIGURE 6 illustrates the change in fiber period caused by stretching and exposing in a clamp an annealed, essentially retracted, fiber of 9.9 per cent N-methylation. The sample exhibited rubbery

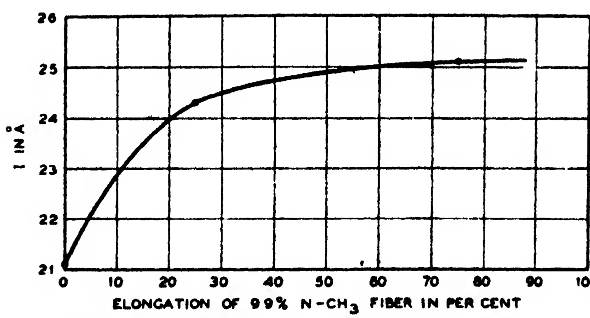


FIGURE 6. Change in identity period, I, caused by stretching fiber of 9.9 per cent N-CH_3 poly decamethylene sebacamide.

[7a] Some evidence for chain retraction also occurs in patterns from dried muscle fibers: see Lotmar, W. & Picken, L. E. R. Helv. Chim. Acta, 25: 538. 1942.

[8] Fuller, C. S., Frosch, C. J., & Pape, N. R. Jour. Am. Chem. Soc. 64: 154. 1942.

behavior throughout this range, and in so far as identity periods reflect distances along a given molecule, FIGURE 6 offers direct evidence of chain lengthening with sample stretch. If rubbers had polar groups which would associate strongly enough to form similarly ordered layers in the retracted state, such changes should likewise be evident in their diagrams.

PLATES 2,A,B and C show the X-ray patterns of the retracted, 25 per cent and 75 per cent stretched fibers, respectively, of 9.9 per cent methylated polymer. The sample for 2,A was not annealed to the complete retraction, hence the equatorial splitting is not well resolved. In 2,B the II_1 layer lines are evident, while in 2,C they have been coalesced again into the meridian. The extended value of I = 25.1 Å from 2,C is, of course, considerably less than the theoretical. However, PLATE 2,D represents the effect of severe stress on even more weakly interacting chains—those with 22.5 per cent methylation. In this pattern, the II_0 features have *d* values of 26.4 Å, and the vestiges of the II_1, of 27.1. These approach the fully extended value of 27.5, which might even be attained with further stretching, but which is, we emphasize again, an unstable configuration. For instance, the unstretched period of this sample was 20.5 Å.

So far, any explanation of the retraction phenomenon would appear related to two factors: the relative proportion of polar to paraffinic matter in the chains, and the strength of interaction in the polar planes. If the chain identity period shortening is from twisting of chain sections about bonds near the polar links, and if such twisting is induced by differences in the preferred packing of the polar and paraffin portions, then weakening of the polar interaction should eventually reduce the observed retraction. This condition apparently occurs when about a third of the amide hydrogens are substituted, in the series studied. For while 22.5 per cent N-methylation readily yields retracted forms of I = 20.5 Å, the 35.6 per cent substituted members give I values of 25.3 Å or more after the usual treatment for retraction. The extensive amorphous portions of the 35.6 per cent polymers confer considerable rubberiness on the samples; the pronounced chain retraction is simply absent from the more ordered regions. PLATE 2,E is the pattern resulting when an oriented filament of 35.6 per cent N-methylation is annealed without tension. For this, and several check samples, I = 23.3 Å. The appearance of meridian spots is partly because of double orientation introduced by stretching ribbons of these soft polymers. Thus, the 2,E pattern is not uniaxial, whereas that of 2,F was established to be so. PLATE 2,F is the pattern obtained after annealing in a clamp, even when

the fiber is unstressed during exposure. Again, I approaches 25.0 Å, and this pattern further illustrates a stable, extended form.

The 54.7 per cent N-methylated polymers show most plainly in this series the rubbery properties of the solids, both mechanically and in diffraction effects. The patterns likewise emphasize the extensive lateral disorder caused when about half of the polar linkages in the dipole layers contain methyl groups. PLATE 2,G was from a sample annealed while clamped, which showed a comparatively high degree of orientation in the stretched condition, but the arcing of 2,G when it was allowed to relax. Accompanying this relaxation is a certain retraction of the identity period, since I = 23.2 Å from the II_0 feature of 2,G. This retraction, however, is not to be considered an equilibrium figure, unlike the retracted values of 20.5 Å. When the sample was stretched 200 per cent and clamped during exposure, the sharpening of PLATE 2,H occurred, and the poorly resolved layer-lines correspond to I = 25.2 Å. Subsequent removal of tension (at room temperature) caused the disorientation of PLATE 2,I. I has returned to 23.2 for the II_0 features, and, of course, the whole sample shortened. Apparently here more than in any previous member of the series, crystallites or ordered regions are orienting and disorienting during the easily reversible rubbery manipulation of the polymer. Correspondingly, and similarly to the phenomenon illustrated in PLATE 2,A,B and C, the sample of 2,G was photographed (not shown) after 100 per cent elongation, and here I = 24.6 Å, intermediate between the 0 and 200 per cent values. The system thus shows a revealing sensitivity to stress. The ordered regions orient and spacings within these regions are expanded, presumably by molecular elongation.

The lateral disorder in the dipole layers which has been predominant in all of the effects thus far discussed may be reviewed in the patterns of PLATE 3,A,B,C,D and E. These are strictly comparable exposures of unoriented sections of compositions 3.5, 9.9, 22.2, 35.6 and 54.7 per cent N-methylation, respectively. Each sample was solidified near its melting point, and represents the highest order characteristic of the compound. The striking effect of the lateral disorder is deterioration of the 3.76 Å reflection, which apparently contributes only diffuse scattering in the 54.7 per cent substituted polymer. The 4.40 Å spacing features remain relatively sharp with increasing N-methylation. So do the inner rings on PLATE 3 that correspond to dipole layer separations along the chains. This is expected since the imperfections are primarily in these layers, and not between them. However, the dissimilar behavior of the

two principal side spacings leads to brief consideration of possible causes of the retraction phenomenon itself.

Retraction Mechanism

Since disorder and intermolecular force weakening in the dipole layers operate in the retraction, we look first for unusual behavior of side-spacing features on the X-ray diagrams. First, the 3.7 Å spacing characteristically splits on the equator and forms layer-line arcs in the retracted forms, as seen in previous figures. Secondly, this same feature deteriorates into diffuse scattering at high degrees of methylation. The other principal side spacing, 4.4 Å, remains intense and virtually unchanged in position through all phases of the extension and retraction. It is, indeed, well resolved in the rubbery sample of 54.7 per cent N-methylation (PLATE 3,E). Such sharpness differs from the scattering found when the entire chains are assumed to be randomly disposed about their long axes.[3,5] The 4.4 Å spacing seems therefore to result from planes whose scattering components are relatively unaffected by disorder in the dipole layers and by the retraction phenomenon. This condition may be attributed to the hydrocarbon sections of the polyamide chains. It may be assumed that these sections tend always to pack like the chains in pure hydrocarbons, whose structures have been previously described.[9] Each hydrocarbon chain section has a polar group, the polymer linkage, at either end. Polar coordination of these groups (association of dipoles and formation of hydrogen bonds) apparently does not favor the paraffin packing. This is reasonable when one imagines the changes in the "setting angle,"[9,10,11] φ, which would occur if attracting polar groups were inserted in the chains shown end-on in FIGURE 7. The chains in the cell drawn would tend to twist at the polar *groups*. Some of the twisting tendency in the dipole layers could occur in the direction of the dotted arrows. There is thus a competition between the preferred positions of the paraffin portions and of the associated polar portions of the chains. Consequently, at each of the dipole layers there is a torque tending to skew portions of the chains from the preceding paraffin arrangements into the polar arrangements, and the result is a rotation about the freer bonds near the polar linkages so that, as seen normal to the fiber axis, the chains progress as illustrated in FIGURE 8(a). Of course, there is doubtless less regularity and uniformity, and certainly less planarity than appear in the scheme of FIGURE 8(a).

[9] Müller, A. Proc. Roy. Soc. 120: 437. 1928.
[10] Kohlhaas, R., & Soremba, K.-H. Zeit. Krist. 100: 47. 1938.
[11] Bunn, C. W. Trans. Faraday Soc. 35: 482. 1939.

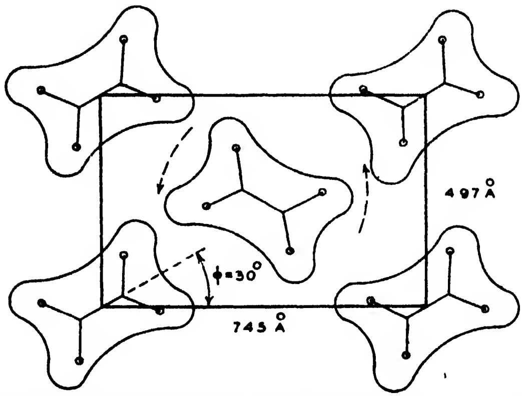


FIGURE 7. Schematic diagram of paraffin chain packing in a plane normal to the chain axes, after Müller. The enclosed areas represent the contours of the force fields.

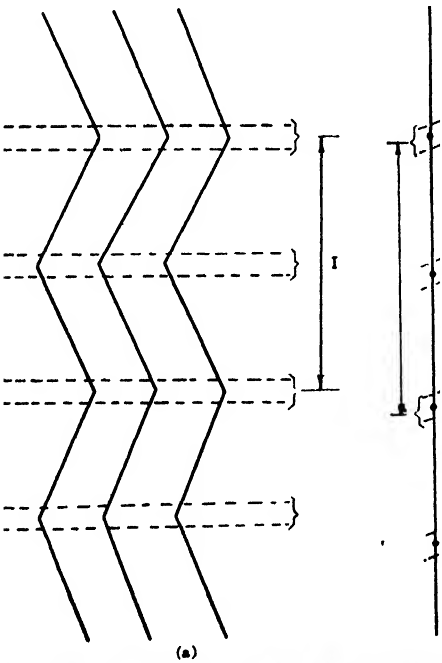


FIGURE 8. Schematic representation of identity period shortening, (a), and of the chains in the fully extended form, (b). The dotted lines represent the boundaries of polar group zones in which the principal interaction and twisting tendency occur.

We may consider briefly the relative energies of the packing competition. Müller [12, 13] has calculated from polarizability and susceptibility values the van der Waals attraction of paraffin chains, which leads to φ values of 0 or 90° for the minimum potential. However, when the repulsion terms are also accounted for, the total potential follows the curve of FIGURE 9, and the minimum is near the experimentally observed angle

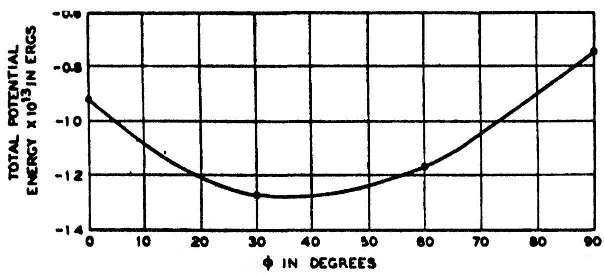


FIGURE 9. Potential energy of interaction of paraffin chains as a function of their "setting angle," φ.

of $\varphi = 30°$. Further, the average energy difference per mole of methylene groups between the 30° and, say, 0° position, can be estimated as 520 calories. This value may be multipled by the number of methylene groups in a paraffin segment in so far as these segments can be regarded as relatively rigid. The energy of rotation around the CH_2–CH_2 bond is apparently high[14] compared to that about such bonds as

$$\overset{\overset{\displaystyle O}{\|}}{CH_2\text{-}C}\,{}^{15},$$

which occurs in the polymers. We may thus approximate the energy of position of the paraffin sections restraining them from occupying the orientation accompanying closest coordination of their polar ends. On the other hand, the coordination energy must be about 6000–8000 calories per hydrogen bond,[7, 16] (two chains). Therefore, a sufficiently high concentration of polar linkages along the chain should yield a structure well dominated by the polar coordination. This results in extended chain configurations as in FIGURE 8(b). These have the

12 Müller, A. Proc. Roy. Soc. 154A: 624. 1936.
13 Müller, A. Proc. Roy. Soc., 178A: 227. 1941.
14 Kistiakowsky, G. B., & Nazmi, F. Jour. Chem. Phys. 6: 18. 1938. Kistiakowsky, G. B., & Rice, W. W. Jour. Chem. Phys. 8: 618. 1940.
15 Shumann, S. C., & Aston, J. G. Jour. Chem. Phys. 6: 485. 1938.
16 Unpublished studies.

calculated length per repeating unit, and well resolved side spacings of 3.70 and 4.40 Å. Polyhexamethylene adipamide is an example of a highly polar structure, with an average of five methylene groups per linkage. When, however, nine methylene groups per linkage obtain, as in polydecamethylene sebacamide, the tendency for a compromise structure is strong, since the energy of position of the hydrocarbon segments is about 4700 calories per mole. The "normal" pattern for high molecular weight or extended 10–10 is like that of PLATE 1,A, but PLATE 3,F is the "compromise," further retracted, form obtained by annealing unclamped a medium molecular weight sample. (The influence of molecular weight in these phenomena is apparently its effect on the internal viscosity, which controls the ease of chain rearrangement.) Likewise, instead of increasing methylene group concentration, or in addition to it, the effect of polar coordination may be lessened by substitution, and then the same results occur. In this case, however, apparently an extreme can be reached at which the compromise structure sufficiently satisfies the preferred hydrocarbon packing so that relatively little twisting and shortening occur, as noted in a previous section. In this condition, only one side spacing remains well defined, however, the 4.4 Å; the 3.7 Å planes have apparently been disordered. Therefore, the chain configurations are actually different from the initial case where extended chains resulted from dipole domination. This disordering is consistent with other behavior of poorly organized systems. For example, only one side spacing (4.18 Å) appears when quenched samples of any polyamide chains are studied. Further, the 3.7 Å planes are those shifted off the equator when the retraction enters. The remaining 4.4 Å planes must be from layers in the plane of FIGURE 8(a), while the 3.8 Å planes are at one or various angles to the plane of the paper. Hence, the whole behavior of the 3.7 Å planes traces the course of the proposed compromise structure.

Thus, the variations in the side-spacing features on the model polymer X-ray diagrams can be related to a general mechanism for chain retraction and extension.

Force Reduction by Solvation

The preceding ideas of chain packing suggest that weakening of polar interaction by solvation should produce similar configuration changes to those caused by substitution or dilution by methylene groups. The solvating agent cannot, of course, be allowed too greatly to disorder the fiber studied. Polyhexamethylene sebacamide, with seven methylene groups per polar linkage, ordinarily shows no trace of a retracted form.

When, however, an annealed fiber was exposed to cresol vapor in equilibrium with the liquid at 120° C. for 12 minutes, the equatorial splitting of the 3.7 Å spacing features and the deterioration of these planes appeared strikingly as in PLATE 3,G. (The untreated pattern closely resembles that of PLATE 1,A.) The changes throughout the pattern are just like those introduced by N-methylation, and are dependent on the actual presence of the cresol. The identity period features from the solvated diagram give I = 17.0 Å, a 24.1 per cent shortening of the theoretical period. A wide variety of hydrogen-bonding agents may be used to cause similar effects, in extent varying with the degree of solvation. Such behavior, incidentally, agrees with the view that specific polar groups in polymers can frequently be solvated without development of extensive disorder in other parts of the same chains. The cresol, like water, bonds at the polar linkages, and small molar amounts of it would, because of its molecular volume, disorder the polar planes. The mechanism of polar plasticizer action is thus illustrated in detail. The fact that such plasticizers can weaken the polar forces without disrupting seriously the paraffin chain packing seems to result in desirable physical properties. One reason is that the paraffin sections of the chain are already plastic enough in the original polymer; the forces between them should not be weakened further. Rather, the polar group forces need modification, and this the polar plasticizer does.

The changes in chain configuration resulting from substitution or solvation represent clear evidence for the importance of *inter*molecular force reduction in permitting chain retraction associated with rubbery elasticity.

Double Orientation

In addition to the influence of lengthwise fiber stress on chain configurations and crystallite structure, the effect of rolling or otherwise compressing uniaxially oriented fibers can be studied. Such treatment results in double orientation, at least of the crystalline regions. The crystallite arrangement in this condition appears to be very similar among all of the even-membered polyamides. From specimens photographed perpendicular to the plane of rolling, meridian spots and equatorial features from the 4.40 Å spacing appear. Sections photographed with the beam in the plane of rolling give only sharply resolved layer-lines and the strong equatorial spots correspond to the 3.70 Å spacings. At about 44 degrees incidence of the beam to the plane of rolling, the intensities of the two equatorial reflections are approximately equal.

PLATE 3,H shows a typical biaxially oriented fiber diagram taken with

the beam perpendicular to the plane of rolling, whereas PLATE 3,I is one with the beam parallel. The sample of these patterns was in the extended form. The crystallites are apparently forced down by the rolling so that the identity period planes (basal planes of the cells) give only meridian reflections when the 3.70 Å planes lie in the plane of rolling (or at nonreflecting angles thereto) and the beam is normal to this plane. In this direction the beam may be imagined to travel along the *Z* axis and strike (*OYZ*) planes, whereas when the beam is parallel to the plane of rolling, it travels along the *Y* axis to strike, figuratively speaking, planes just displaced by the glancing angle from the (*XOZ*) plane, and layer-line reflections are thereby produced.

Higher orientation of the retracted forms of the fibers would have yielded desirable information on their crystallite order. However, repeated experiments always indicated that the retracted form was largely transformed to the extended by the cold-working accompanying rolling or other necessary types of compression. It may be noted the physical properties such as elastic modulus discussed previously represent chiefly the retracted form, and somewhat different values might be expected from the extended state. A comparison of the melting points of the two forms would be especially interesting, since the unstable extended form is presumably retained at melting temperatures only by tension. This is reminiscent of the melting range in stretched rubber, for which varying local tensions are sometimes held responsible.

Chain Packing in Copolyamides

The extensive interchain force reduction attending the dipole layer disorder in copolyamides[4] might be expected to provide conditions for retracted chain forms. These were always found to some extent, although incipient retracted type patterns occurred only in a few instances, such as for the 50 per cent copolymer polyhexamethylene sebacamide-polydecamethylene adipamide (6–10:10–6), and others containing the often retracted 10–6 units. This poor definition of the retracted form in the copolyamides as contrasted to the N-substituted series agrees with the idea that not only polar force weakening but also the strong interaction of the hydrocarbon portions is required for extensive retraction. In other words, the copolyamides do not contain sufficient competition of the hydrocarbon chain and dipole packing forces. For in the copolyamides (especially in the middle percentages) a great proportion of the polar linkages are so displaced that they interfere with the orderly packing tendencies of the methylene chains, which might otherwise tend to twist into the shortened form. Copolymeriza-

tion may thus be regarded as more complex than substitution. In copolymers, not only are the polar layers disorganized, but also the packing of the chain sections between them is disrupted by occasional occurrence of groups and linkages which belong in the dipole regions. This difference is useful in analyzing the behavior of many polymers, including numerous vinyl derivatives. Cellulose derivatives, on the other hand, are evidently capable in general of showing only the behavior characteristic of substituted polymers.

The chain disruption noted above, caused by the displaced linkages in copolyamides, may be illustrated in another way. The systematic relation of elastic modulus to melting point has been shown for the substituted series, and a similar comparison has been discussed for the copolymers.[4] Comparison of the two general types shows a different result. For example, the copolyamide from 66 per cent of 6–6 (polyhexamethylene adipamide) and 33 per cent of 10–10 (polydecamethylene sebacamide) base components (reacted at random) melts at 174° C., 9.9 per cent N-methylated 10–10 melts at 175° C. These values may be considered the same, and are at about the limiting values (convergence temperatures) for increase of melting point with chain length. The Young's modulus of the copolyamide (1.2×10^9 dynes/cm^2) is 20 per cent higher than that of the N-substituted polymer (1.0×10^9) of the same melting point. This is to be expected, since the copolyamide actually contains a greater concentration of interacting polar groups, despite its disorder, than the substituted one and is thus harder. The significant factor seems to be that the displacement of the polar groups so disturbs the chain packing that the copolymer, in spite of its higher polar group concentration, shows the same melting point as the substituted compound.

The authors wish to thank Dr. B. S. Biggs and Mr. W. S. Bishop for supplying the polyamides, and Messrs. N. R. Pape and J. H. Heiss, Jr., for considerable help in connection with the experimental work. They also wish to express their gratitude to the Editor of the Journal of the American Chemical Society for permission to use material previously submitted to the Journal.

SUMMARY

A series of 9 N-methylated polyamides, with methyl substitution varying in polydecamethylene sebacamide from 0 to 55 mol per cent, has been studied as a representative group of linear polymers showing physical properties ranging from hard brittleness to rubberiness. Young's modulus, moisture sorption and relative solubility were chosen

as properties representing the gross solids, while the corresponding fine structure was studied by X-ray diffraction from oriented and unoriented sections.

The elastic modulus and hardness decrease rapidly with increasing N-methylation, as the hydrogen bonding and other polar forces decline. Relative solubility increases. Moisture sorption also increases, since the disorder introduced by the N-methylation leaves polar groups uncoordinated, and hence free to sorb water. This is in spite of the somewhat hydrophobic character of the methyl group, which replaces bonding hydrogen.

The interchain spacings are not appreciably changed by the methyl substitution, but at higher amounts of substitution one of the principal spacings (3.76 Å) becomes diffuse. However, the spacings between the dipole layers, related to the identity period along the chain, are more than 25 per cent shorter in the crystal form characteristic of the lower ranges of N-methylation than in the normal extended form. The chains appear to be retracted by partial folding along the fiber axis. This retracted form can be converted to an extended one by stretching in which the fiber period lengthens (as from 20.5 to 26.5 Å) and the structure changes also. This behavior is reminiscent of fibrous proteins, as are many other properties of these polymers.

Similar retraction phenomena have been introduced in normally extended, unsubstituted polyamides by allowing plasticizers, such as cresol, to penetrate into the polar layers. The resulting structural changes demonstrate polar association of the plasticizer, often considered as a mechanism for plasticizer action.

A possible explanation of chain twisting is that it results from a compromise of the packing tendencies of the paraffin sections and polar sections of the chain. In general, it may reflect competing packing tendencies.

The chain retraction and extension observed may be the first stages of rubbery elasticity.

EXPLANATION OF FIGURES

Plate 1

X-ray patterns (where oriented, with fiber axis vertical) of *N*-methylated polydecamethylene sebacamides. The number refers to the mol per cent of $N-CH_3$ groups.

A. Extended form, 3.5 per cent.
B. Retracted form, 3.5 per cent.
C. Quenched fiber, 9.9 per cent.
D. Low temperature annealing, 5.7 per cent.
E. Sample of (D), stretched.
F. Quenched, unoriented, 9.9 per cent.
G. Sample of (F), annealed in solid state.
H. Cooled slowly from melt, unoriented, 9.9 per cent.

Plate 2

X-ray fiber patterns (axis vertical) of N-methylated polydecamethylene sebacamides showing effects of tension.

A. Fiber of 9.9 per cent $N-CH_3$, annealed two hours at 162° C., unclamped.
B. Sample of (A), stretched 25 per cent.
C. Sample of (A), stretched 75 per cent.
D. Fiber of 22.5 per cent $N-CH_3$, stretched to maximum.
E. Fiber of 35.6 per cent $N-CH_3$, annealed, unclamped.
F. Sample of (E), annealed in clamp, exposed unclamped.
G. Fiber of 54.7 per cent $N-CH_3$, annealed in clamp, exposed unclamped.
H. Sample of (G), stretched 200 per cent, exposed clamped
I. Sample of (H), exposed unclamped.

Plate 3

X-ray patterns of variously-ordered polyamides.

A–E. Debye-Scherrer diagrams of slowly solidified samples showing maximum crystallinity (percentages of N-methylation shown): A, 3.5 per cent; B, 9.9 per cent; C, 22.2 per cent; D, 35.7 per cent; E, 54.7 per cent.

F. Retracted form of low molecular weight unsubstituted polydecamethylene sebacamide.

G. Cresol-treated fiber of polyhexamethylene sebacamide.

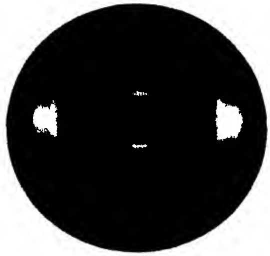
H. Biaxially-oriented 9.9 per cent N-methylated polydecamethylene sebacamide, beam perpendicular to plane of rolling.

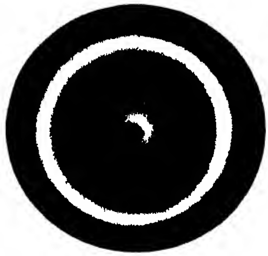
I. Biaxially-oriented sample with beam parallel to plane of rolling.

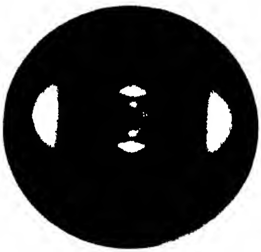
EXPLANATION OF PLATES

X-ray patterns of linear polymers, showing the effects on molecular structure produced by stretching, quenching, annealing and other types of treatment that modify interchain forces.

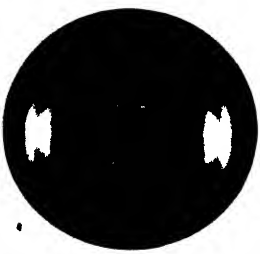
A







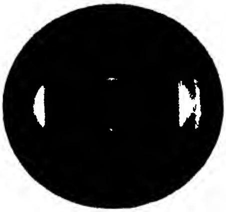
C



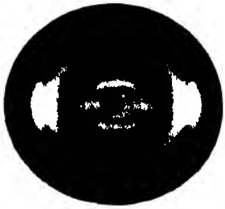
D

H

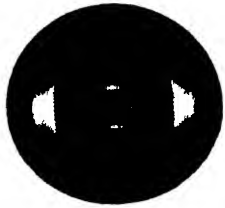
BAKER AND FULLER. LINEAR POLYMERS



A

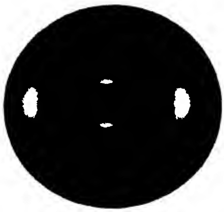


B



C

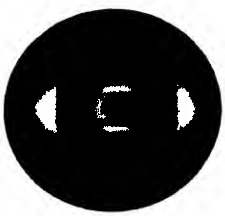
D



E

F

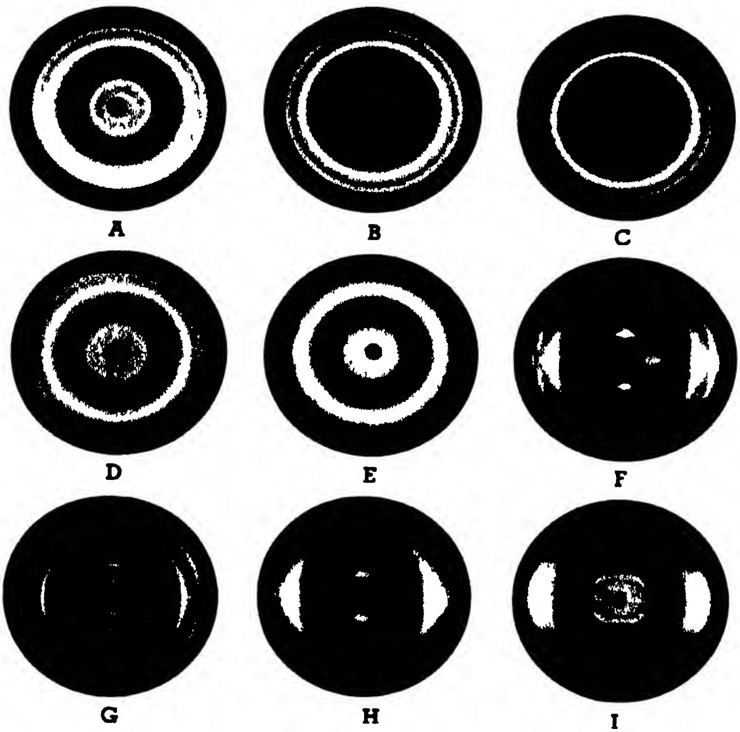
G



H

I

BAKER AND FULLER LINEAR POLYMERS



BAKER AND FULLER LINEAR POLYMERS

SOME ASPECTS OF THE MECHANISM OF ADDITION POLYMERIZATION

By Charles C. Price
From the Noyes Chemical Laboratory, University of Illinois, Urbana, Illinois

INTRODUCTION

It has long been recognized and is now generally accepted that, as a rule, the process of addition polymerization differs fundamentally from condensation polymerization. The latter involves only types of reactions with which the chemist is quite familiar, with the important characteristic feature that the monomer molecules are so constructed that the condensation process is capable of indefinite repetition to build up a polymeric product. The reaction proceeds in a regular, stepwise manner, yielding material of steadily increasing degree of polymerization throughout the course of the reaction.

This mode of formation contrasts sharply with that involved in typical addition polymerizations, which are characterized by the fact that the polymer molecules first formed are essentially the same size as the rest. In other words, the polymer molecules do not grow gradually, but each polymer molecule grows rapidly to a certain size and is then stabilized. This behavior, coupled with the susceptibility of these reactions to catalysis and inhibition, supports the suggestion that these reactions are chain reactions involving activation of the monomer (*initiation*) followed by rapid successive additions of monomer molecules to the active intermediate (*propagation*) until the activated polymer becomes stabilized (*cessation*).

A. Initiation

$$M \rightarrow M^*$$

B. Propagation

$$M^* + M \rightarrow M_2^* \rightarrow \rightarrow \rightarrow M_n^*$$

C. Cessation

$$M_n^* \rightarrow M_n$$

In addition to the cessation reaction, leading to destruction of the

active center, indicated above by an asterisk, it has been suggested that under certain conditions the activity could be transferred to some other molecule, monomer, polymer or solvent.[1]

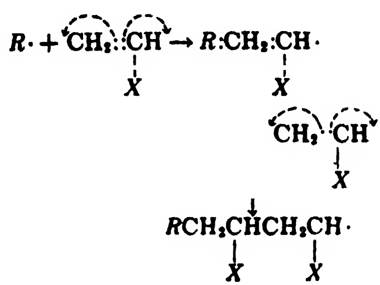
D. Chain Transfer

$$M_n^* + X \rightarrow M_n + X^*$$

This general outline for the course of addition polymerization has been very generally accepted. Major points of speculation and controversy, however, have been the specific nature of the activation, indicated above simply by an asterisk, and the manner in which this activation is accomplished. It seems apparent now that one source of confusion has been the fact that the activation may be accomplished in many different ways and may even be of a different character for polymerizations carried out under different conditions or with different monomers.

CHEMICAL EVIDENCE FOR FREE RADICAL ACTIVATION

Staudinger[2] was one of the first to suggest that this activation might consist of a free radical.



Such a process obviously will result in the formation of a polymeric product in which the monomer units will be joined by covalent bonds. This suggestion has been convincingly supported by the fact that a large

[1] Flory, P. J. Jour. Am. Chem. Soc. 59: 241. 1937.
[2] Staudinger, H. "Die hochmolekularen organischen Verbindungen." Julius Springer. Berlin. 1932.

variety of reactions which, on the basis of entirely independent evidence,[3] are believed to generate active free radicals ($R\cdot$ above) have been shown to initiate many addition polymerizations. Some of these are indicated below.[4–8]

$$R-\overset{\displaystyle O}{\overset{\|}{C}}-O-O-\overset{\displaystyle O}{\overset{\|}{C}}-R \rightleftharpoons 2R\overset{\displaystyle O}{\overset{\|}{C}}-O\cdot \rightarrow R\cdot + CO_2$$

$$(C_6H_5)_2\underset{}{\overset{\displaystyle CN}{\overset{|}{C}}}-\underset{\underset{\displaystyle CN}{|}}{C}(C_6H_5)_2 \rightleftharpoons 2(C_6H_5)_2\underset{\underset{\displaystyle CN}{|}}{C}\cdot$$

$$CH_3\overset{\displaystyle O}{\overset{\|}{C}}-H \xrightarrow{h\nu} CH_3\cdot + H\overset{\displaystyle O}{\overset{\|}{C}}\cdot$$

$$H_2 \underset{Hg}{\xrightarrow{h\nu}} 2H\cdot$$

$$Ar-N=N-OH \rightarrow Ar\cdot + N_2 + HO$$

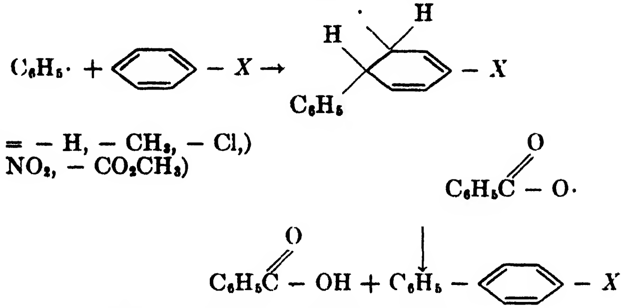
$$\left.\begin{array}{l} Ar-N=N-OCOCH_3 \\ Ar-N\begin{array}{l}\diagup COCH_3 \\ \diagdown N=O\end{array} \end{array}\right\} \rightarrow Ar\cdot + N_2 + CH_3CO_2\cdot \rightarrow CH_3\cdot + CO_2$$

Some of the most convincing evidence for decomposition of peroxides, diazohydroxides, and nitrosoacetanilides to generate active free radicals is the coupling principally in the *para*-position on reaction with benzene derivatives, regardless of the orienting influence of the substituent, and the reaction with carbon tetrachloride to form hexachloroethane.[3]

[3] Hey, M. H., & Waters, V. F. Chem. Rev. 21: 169. 1937.
[4] Schulz, G. V., & Wittig, G. Naturwiss. 27: 387, 456. 1939
[5] Schulz, G. V. Naturwiss. 27: 659. 1939.
[6] Melville, W. H. Proc. Roy. Soc. (London) A163: 511. 1937.
[7] Melville, W. H. Trans. Inst. Rubber Ind. 15: 209. 1939
[8] Price, C. C. & Durham, D. A. Jour. Am. Chem. Soc 64: 2508 1942.

$$C_6H_5\overset{O}{\overset{\|}{C}} - O - O - \overset{O}{\overset{\|}{C}} - C_6H_5 \rightleftarrows 2C_6H_5\overset{O}{\overset{\|}{C}} - O\cdot$$

$$C_6H_5\overset{O}{\overset{\|}{C}} - O \rightarrow C_6H_5\cdot + CO_2$$



$(X = -H, -CH_3, -Cl,)$
$(-NO_2, -CO_2CH_3)$

$$C_6H_5\cdot + CCl_4 \rightarrow C_6H_5Cl + CCl_3\cdot$$

$$2CCl_3\cdot \rightarrow C_2Cl_6$$

The only logical and reasonable explanation for the lack of orienting influence in the substitution in aromatic compounds is that the entering radical is neutral and therefore not subject to orientation due to polarization of the aromatic ring. The ease of free radical substitution in aromatic nitro compounds and quinones[9,10] may be accounted for by increased resonance stability of the intermediate addition compound.[11]

Propagation by reaction of such free radical fragments with the monomer would lead to polymer containing fragments from the catalyst as end groups.

$$(RCO_2)_2 \rightleftarrows 2RCO_2 \rightarrow R + CO_2$$

$$R\cdot + CH_2 = \underset{X}{\underset{|}{CH}} \rightarrow RCH_2\underset{X}{\underset{|}{CH}} \rightarrow\rightarrow\rightarrow$$

$$R\overset{O}{\overset{\|}{C}} - O + CH_2 = \underset{X}{\underset{|}{CH}} \rightarrow R\overset{O}{\overset{\|}{C}} - OCH_2\underset{X}{\underset{|}{CH}} \rightarrow\rightarrow\rightarrow$$

[9] Fieser, L. F., & Oxford, A. E. Jour. Am. Chem. Soc. 64: 2060. 1942
[10] Fieser, L. F., Clapp, R. C., & Daudt, W. H. Jour. Am. Chem. Soc. 64. 2052. 1942
[11] Price, C. C., & Durham, D. A. Jour. Am. Chem. Soc. 65: 757 1943

We have prepared more than twenty samples of polystyrene and polymethyl methacrylate in the presence of anisoyl,[12] *p*-bromobenzoyl,[12] 3,4,5-tribromobenzoyl,[13] and chloroacetyl peroxides[12] and *p*-bromobenzene-diazohydroxide[8] and each sample has contained fragments from the peroxide.

A similar observation has been made for catalysis by tetrachlorotetraphenylsuccinonitrile.[14] Correlation of the analytical data with the molecular weight determinations has indicated an average of very close to one catalyst fragment per polymer molecule for the low molecular weight polymers investigated. A few examples follow.

$$C_6H_5CH{=}CH_2 \xrightarrow[\substack{C_6H_6\\80^\circ}]{(BrC_6H_4CO_2)_2} \underset{(80\%)}{BrC_6H_4(C_8H_8)_{12}}$$

Calculated: C, 87.14; H, 7.17; Br, 5.69; mol. wt., 1400. Found: C, 86.95; H, 7.41; Br, 5.39, 5.72; mol. wt., 1600 (visc.).[15]

$$C_6H_5CH{=}CH_2 \xrightarrow[\substack{C_6H_6\\80^\circ}]{(Br_3C_6H_2CO_2)_2} Br_3C_6H_2(C_8H_8)_{12}O_2$$

Calculated: C, 76.02; H, 6.13; Br, 14.87; mol. wt., 1600. Found: C, 76.27; H, 6.15; Br, 14.40, 15.05; mol. wt., 1710, 1880, 2390 (cryoscopic); 1430 (visc.).[15]

$$C_6H_5CH{=}CH_2 \xrightarrow[95^\circ]{(ClCH_2CO_2)_2} ClCH_2(C_8H_8)_{45}$$

Calculated: C, 91.55; H, 7.70; Cl, 0.75; mol. wt., 4730. Found: C, 91.54; H, 7.53; Cl, 1.05; mol. wt., 5000 (visc.).[15]

$$\begin{array}{c} CH_3 \\ | \\ C = CH_2 \\ | \\ CO_2CH_3 \end{array} \xrightarrow[\substack{\text{dioxane}\\25^\circ}]{(ClCH_2CO_2)_2} ClCH_2(C_5H_8O_2)_{80}O$$

Calculated: C, 59.65; H, 8.02; Cl, 0.439; mol. wt., 8075. Found: C, 59.68; H, 8.38; Cl, 0.396; mol. wt., 8050 (visc.).

$$C_6H_5CH{=}CH_2 \xrightarrow[0^\circ]{BrC_6H_4N_2OH} BrC_6H_4(C_8H_8)_{17}OH$$

---

[12] Price, C. C., Kell, R. W., & Krebs, E. Jour. Am. Chem. Soc. 64: 1103. 1942.
[13] Price, C. C., & Tate, B. E. Jour. Am. Chem. Soc. 65: 517. 1943.
[14] See Mark, H. Ann. N. Y. Acad. Sci. 44: 267. 1943.
[15] The viscometric estimates of molecular weight of polystyrene were evaluated by using Kemp, R. A., & Peters, H. (Ind. Eng. Chem. 34: 1097. 1942) value for the constant in Staudinger's viscosity equation. This has given values in rough agreement with our cryoscopic measurements.

Calculated: C, 87.75; H, 7.31; Br, 4.11; mol. wt., 1950. Found: C, 87.77; H, 7.24; Br, 4.2; mol. wt., 2300 (visc.).[15]

For a few samples above, the absence of oxygen indicates that the fragments from the peroxide must have been almost exclusively alkyl radicals rather than the corresponding acyloxy radicals. The isolation of *p*-bromobenzoic acid from saponification of a sample of polystyrene prepared from *p*-bromobenzoyl peroxide is, however, convincing evidence that under some conditions the acyloxy radicals may also be present in considerable or even preponderant proportion.[16]

KINETIC EVIDENCE FOR FREE RADICAL PROPAGATION

Several investigations of the rate of peroxide-catalyzed polymerization of styrene[17] and vinyl acetate[18] using high concentrations of monomer have indicated a rather complex dependence of rate on monomer concentration. The interpretation of these data, however, is complicated by the difficulty of the unknown effect of the changing solvent medium on the kinetics.[19] The rates of the benzoyl peroxide-catalyzed polymerization of vinyl *l*-β-phenylbutyrate and *d*-*s*-butyl α-chloroacrylate in dilute solution gave rates very accurately first-order with respect to the monomer,[20] as well as half-order with respect to the catalyst.[21]

Investigation of peroxide-catalyzed polymerizations of styrene,[17] methyl methacrylate,[22] *d*-*s*-butyl α-chloroacrylate[21] and vinyl acetate[18] have all clearly indicated a rate dependent on the square root of the catalyst concentration. On the basis of an initiation process first-order with respect to the catalyst, this evidence has been universally interpreted as conclusive evidence for a cessation process second-order with respect to active centers. This offers substantiation for the free radical mechanism since, in the absence of inhibitors, destruction of free radical activity can only occur by reaction of *two* active radicals. In other words, the activity of the free radical due to the odd electron can only be destroyed by pairing with the odd electron of a second radical.

The data indicating the accuracy with which the benzoyl peroxide-catalyzed polymerization of *d*-*s*-butyl α-chloroacrylate follows kinetics first-order with respect to monomer and half-order with respect to catalyst[21] are summarized in the accompanying figures. These data can be readily interpreted on the basis of a first-order decomposition of

---

[16] Bartlett, P. D., & Cohen, S. G. Jour. Am. Chem. Soc. 65: 543. 1943.
[17] Schulz, G. V., & Husemann, E. Zeit. physikal. Chem. B39: 246. 1938.
[18] Medvedev, S., & Kamenskaya, S. Acta Physicochim. USSR 13: 565. 1940.
[19] Breitenbach, J. W. Zeit. physikal. Chem. B45: 101. 1939.
[20] Marvel, C. S., Dec, J., & Cooke, H. G. Jour. Am. Chem. Soc. 62: 3499. 1940.
[21] Price, C. C., & Kell, R. W. Jour. Am. Chem. Soc. 63: 2798. 1941.
[22] Norrish, R. G. W., & Brookman, E. F. Proc. Roy. Soc. (London) 171A: 147. 1939.

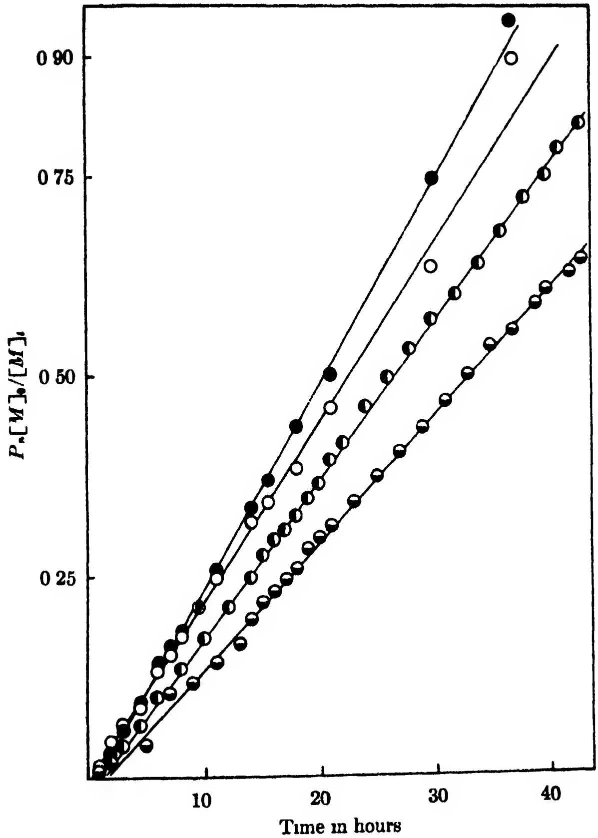


FIGURE 1 Rate of benzoyl peroxide-catalyzed polymerization of *d*-*s* butyl α-chloroacrylate at 26° C *

	[M](g /100 cc)	[Cat](g /100 cc)	k'	$k(=k'/[Cat]^{1/2})$
◐	5 63	2 00	0 0194	0 0137
◒	5 82	1 67	0155	0120
○	7 09	2 67	0229	0140
●	4 15	3 20	0257	0143
				0 0135 (±6%

* Joseph Dec, Ph D Thesis, University of Illinois, June, 1940

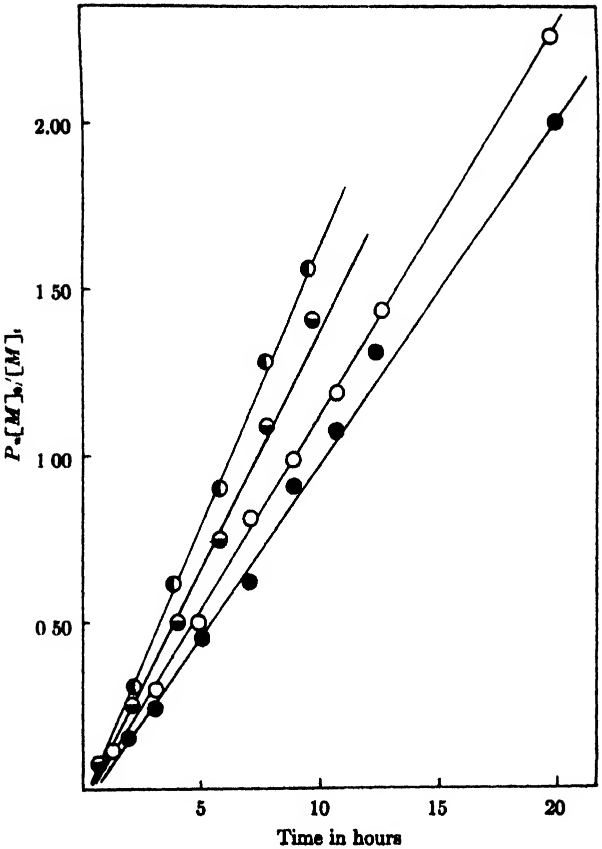


FIGURE 2. Rate of benzoyl peroxide-catalysed polymerization of *d-s*-butyl α-chloroacrylate at 44° C.

	[M](g./100 cc.)	[Cat](g./100 cc.)	k'	$k(=k'/[Cat]^{1/2})$
○	8.74	4.58	0.112	0.0523
●	5.33	4.01	.100	.0515
◐	3.64	11.22	.163	.0498
⊖	3.64	7.50	.138	.0514
				0.0515 (±2%)

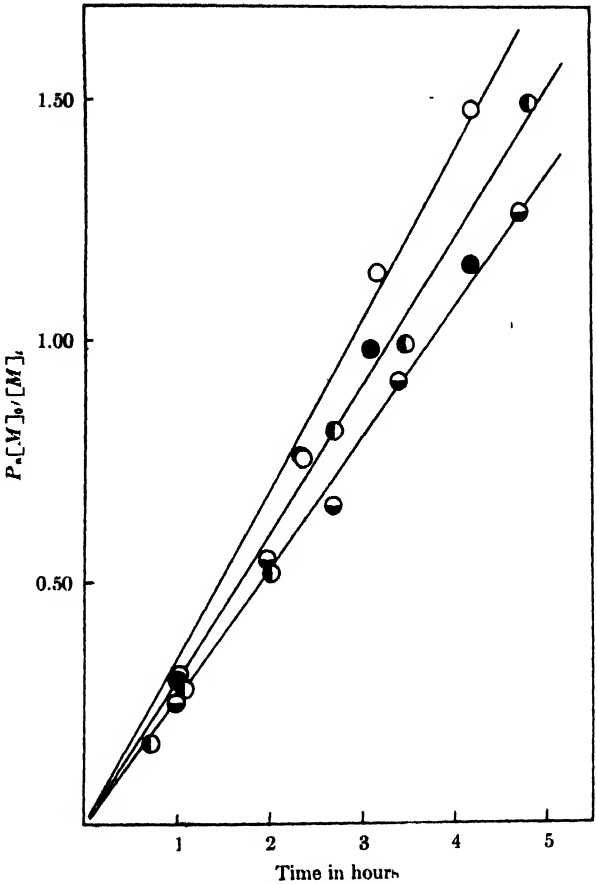


FIGURE 8. Rate of benzoyl peroxide-catalyzed polymerization of *d*-*s*-butyl α-chloroacrylate at 52° C.

	[M](g./100 cc.)	[Cat](g./100 cc.)	k'	$k (= k'/[\mathrm{Cat}]^{1/2})$
○	8.64	11.22	0.341	0.102
●	8.64	7.50	.200	.101
◐	8.64	7.50	.300	.101
	8.64	5.62	.268	.113
				0.104 (± 4%)

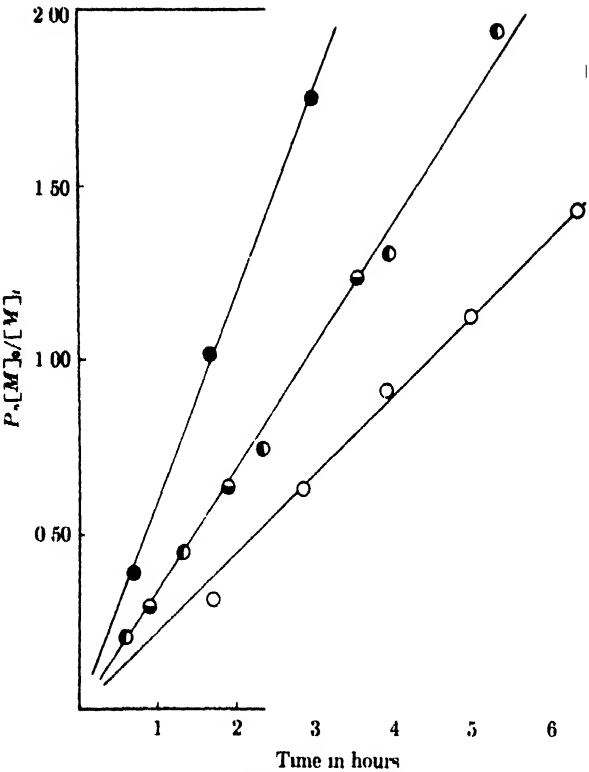


FIGURE 4 Rate of benzoyl peroxide-catalyzed polymerization of *d*-*s*-butyl α-chloroacrylate at 60° C

	[M](g /100 cc)	[Cat](g /100 cc)	k'	$k(=k'/[Cat]^{1/2})$
○	4 98	2 29	0 230	0 157
◐	4 98	4 58	342	160
◒	4 98	4 58	342	160
●	4 98	9 16	569	185

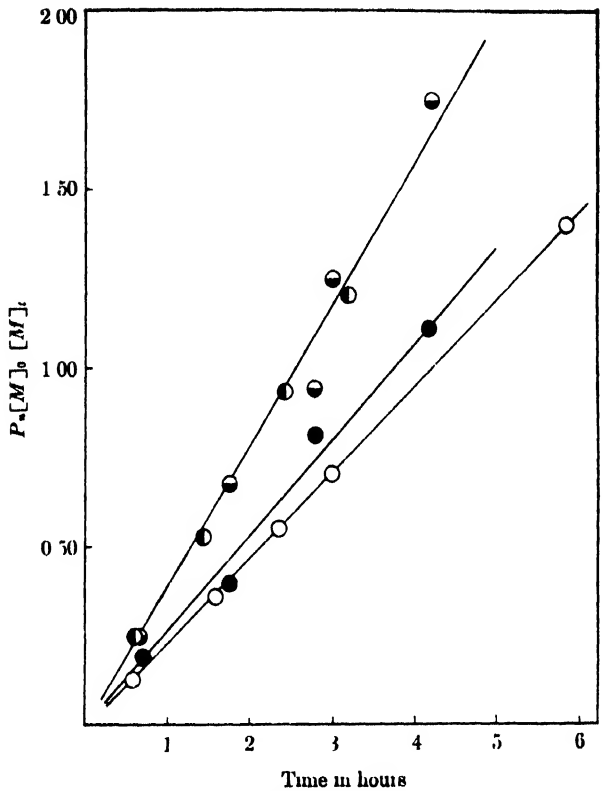


FIGURE 5 Rate of benzoyl peroxide-catalyzed polymerization of *d*-*s*-butyl α-chloroacrylate at 60° C

	[M](g./100 cc.)	[Cat.](g./100 cc.)	k'	$k(=k'/[\text{Cat}]^{1/2})$
○	9 91	2 29	0 244	0 162
◐	9 91	4 58	380	177
◒	2 48	4 58	380	178
●	2 48	2 28	268	175
				0 169 (±6%)

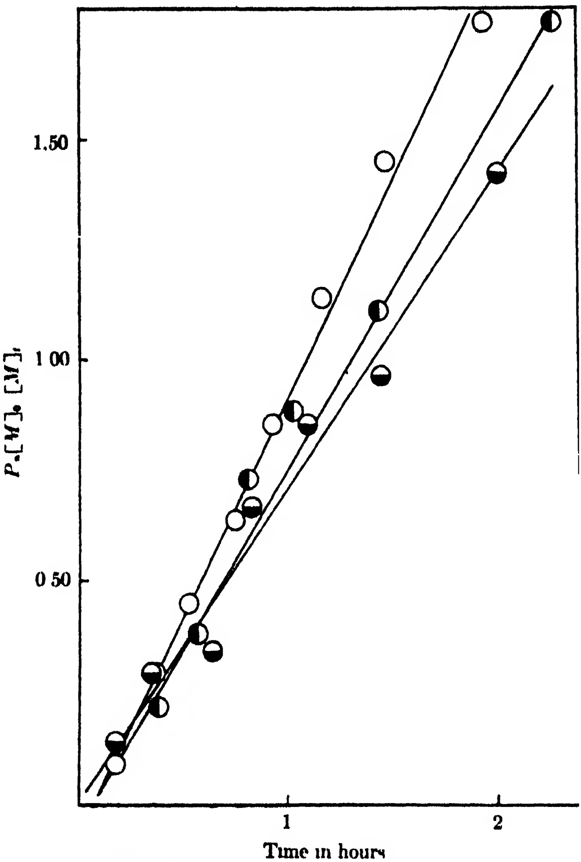


FIGURE 6 Rate of benzoyl peroxide-catalyzed polymerization of *d-s*-butyl α-chloroacrylate at 68° C

	[M](g./100 cc.)	[Cat](g./100 cc.)	k'	$k(=k'/[\mathrm{Cat}]^{1/2})$
○	7.88	7.50	0.936	0.342
◐	11.43	7.50	.795	291
◒	11.43	5.62	.714	301
				0.311 (±7%)

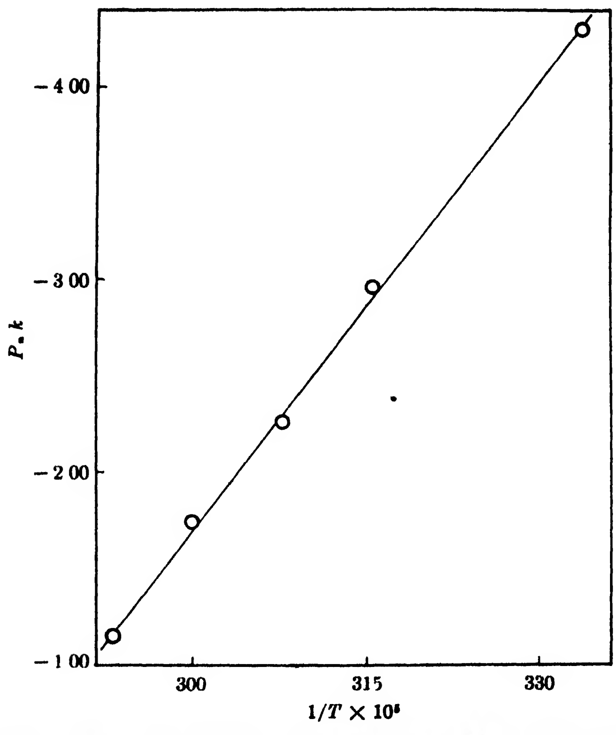


FIGURE 7 Temperature dependence of benzoyl peroxide-catalyzed polymerization of *d-s*-butyl α-chloroacrylate

Over-all "Activation Energy" + 15,200 ± 400 cal

benzoyl peroxide to form the initiating active radicals, a second-order propagation by reaction of active radicals with monomer (M) and a second-order cessation by reaction of two active radicals.

A. Initiation

$$(C_6H_5CO_2)_2 \xrightarrow{k_a} \underbrace{C_6H_5CO_2\cdot + C_6H_5\cdot}_{R\cdot} + CO_2$$

B. Propagation

$$R\cdot + M \xrightarrow{k_b} RM\cdot \xrightarrow{k_b}\xrightarrow{k_b}\xrightarrow{k_b} RM_n\cdot$$

C. Cessation

1. $RM_n\cdot + RM_m\cdot \xrightarrow{k_{c1}} RM_n + RM_m$ (or $RM_{(n+m)}R$)

2. $RM_n\cdot + R\cdot \xrightarrow{k_{c2}} RM_nR$

3. $2C_6H_5CO_2\cdot \xrightarrow{k_{c3}} (C_6H_5CO_2)_2$

The rate of disappearance of monomer is represented by the expression $-d[M]/dt = k_b[RM_x\cdot][M]$, where x may vary from zero to n. If the rate constants of the reactions of the active radicals (k_b and k_c) are large compared with the rate constant of their formation (k_a), a steady state with respect to the rate of formation and destruction of active centers in the system will be rapidly reached.

$$d[RM_x\cdot]/dt = k_a[\text{Cat}] = -d[RM_x\cdot]/dt = k_c[RM_x\cdot]^2$$
$$[RM_x\cdot] = \sqrt{k_a[\text{Cat}]/k_c}$$

Substitution of this value for the concentration of the active intermediates in the equation for the rate of disappearance of monomer gives the experimentally observed dependence of the rate on both monomer and catalyst concentrations.

$$-d[M]/dt = k'[\text{Cat}]^{\frac{1}{2}}[M]$$

It should be clearly recognized that, on the basis of this treatment, there is no "rate-controlling" step in the usual sense of this term. The rate constant for the over-all polymerization process (k' above) is a composite of the rate constants of all three steps in the process ($k' = k_b k_a^{\frac{1}{2}}/k_c^{\frac{1}{2}}$). Therefore, the effect of temperature on k' does *not* give directly an activation energy for any one step in the process.

It should also be pointed out that, in the strict sense of the word, the peroxide is not a catalyst, since it is not regenerated at the end of the process and, in fact, fragments from it are chemically bound to the polymer.

Several reports of the kinetics of the decomposition of benzoyl peroxide have indicated that this reaction is actually a first-order process as indicated above.[23, 24] However, during the benzoyl peroxide-catalyzed polymerization of vinyl acetate[18] and of styrene[16] and the 3,4,5-tribromobenzoyl peroxide-catalyzed polymerization of styrene,[13] the rate of disappearance of peroxide is increased several fold.[13] A possible explanation may be that in thermal decomposition, only that portion of the decomposition which leads to an alkyl radical and carbon dioxide is measured since the acyloxy radicals may recombine to form the peroxide. In the presence of a monomer which will react with the acyloxy as well as the alkyl radicals, reversion to peroxide will be inhibited and the over-all rate of disappearance of peroxide will be increased.

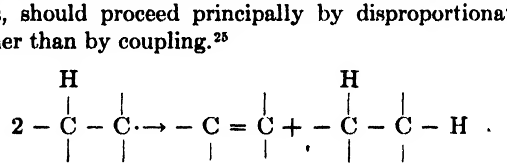
Further evidence for the second-order cessation reaction may be derived from the fact that the degree of polymerization is inversely proportional to the square root of the peroxide catalyst concentration,[17] which is readily interpreted on the basis of a propagation first-order and a cessation second-order with respect to the active radicals.

$$\text{Chain length} = \frac{\text{rate of propagation}}{\text{rate of cessation}} = \frac{k_b[M][RM_x\cdot]}{k_c[RM_x\cdot]^2} = k'' \frac{[M]}{[\text{Cat}]^{1/2}}$$

The chain length here refers to the kinetic chain length of the reaction as a chain reaction and is equivalent to the degree of polymerization *only* if there is no chain transfer.

FREE RADICAL CESSATION. CHAIN TRANSFER AND INHIBITION

Studies of both the rate and the degree of polymerization, as mentioned above, have produced convincing evidence that the cessation reaction for free radical polymerization is bimolecular. It has frequently been suggested that such a process as this, involving the reaction of two free radicals, should proceed principally by disproportionation when possible rather than by coupling.[25]



The observations that a number of low molecular weight polymer molecules contained one fragment per polymer molecule seem to substantiate

[23] Brown, D. J. Jour. Am. Chem. Soc. 62: 2657. 1940.
[24] McClure, J. H., Robertson, R. E., & Cuthbertson, A. C. Can. Jour. Res. 20B: 103. 1942.
[25] Rice, F. O., & Rice, K. K. "The Aliphatic Free Radicals," Johns Hopkins Press Baltimore 1935.

disproportionation. It is perhaps pertinent to point out that the coupling of phenyl, triphenylmethyl, methyl, trichloromethyl and acyloxy radicals is not in disagreement with this idea since in these instances disproportionation in the sense indicated above is not possible.

In addition to stabilization of active growing polymer molecules by cessation, they may also be stabilized by chain transfer. For example, the observation that each molecule of polystyrene prepared in the presence of carbon tetrachloride contained the elements of carbon tetrachloride[26] may be readily explained by a chain transfer process involving the known reaction of active free radicals with carbon tetrachloride.[3]

$$M_n\cdot + CCl_4 \rightarrow M_nCl + CCl_3\cdot$$

$$\begin{array}{l} CCl_3\cdot + M \rightarrow Cl_3CM\cdot \rightarrow \rightarrow \rightarrow Cl_3CM_n\cdot \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad {\scriptstyle CCl_4}\big\downarrow \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad Cl_3CM_nCl + CCl_3\cdot \end{array}$$

In the discussion of this paper, Dr. Bartlett pointed out that the proportion of chain transfer process involving monomer would increase markedly as one changed from conditions for forming low molecular weight to those for forming high molecular weight polymers.

$$R\text{-}(M)_x\text{—}CH_2\underset{\underset{X}{|}}{CH}\cdot + CH_2 = \underset{\underset{X}{|}}{CH} \rightarrow R\text{-}(M)_x\text{—}CH = \underset{\underset{X}{|}}{CH} + CH_3\underset{\underset{X}{|}}{CH}\cdot$$

$$\downarrow\downarrow\downarrow$$

$$H(M)_x\cdot$$

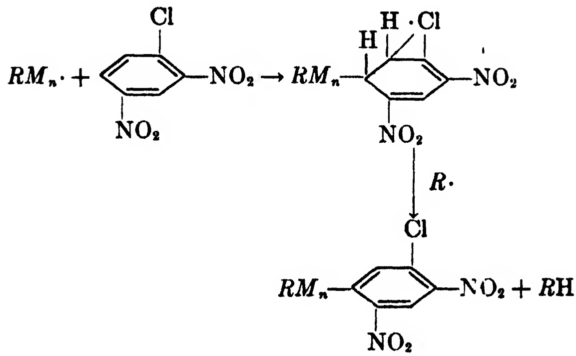
The chance of interrupting the growth of any particular active chain by cessation will vary with the square of the concentration of active chains, while the chance for the above chain transfer process will be directly proportional to the concentration of active chains. Dr. Bartlett therefore predicted that, for very high molecular weight polymers, it may be found that there will be considerably less than one catalyst fragment per polymer molecule. Dr. Mark and Dr. Flory both suggested that there are already certain experimental observations in agreement with this view.

On the basis of the active free radical mechanism for the polymerization, an inhibitor or retarder for the reaction will be any substance which will destroy this activity. Unless the inhibitor is itself a free radical (triphenylmethyl) it cannot, of course, destroy the free radical. Rather,

[26] Breitenbach, J. W., & Maschin, H. Zeit physikal Chem. A187. 175 1940

it must react with the "active" free radical to give an "inactive" free radical, where "active" and "inactive" refer only to whether or not the radicals are reactive enough to add to a monomer molecule.

The retardation of polymerization by aromatic nitrocompounds can be very reasonably accounted for on this basis. Since nitromethane does not retard the polymerization of styrene,[11] and since it has been demonstrated that nitro groups strongly activate the aromatic nucleus toward free radical substitution,[10] the logical explanation for the retarding effect of aromatic nitrocompounds must be their reaction with the growing radical chain to give a *relatively* stable radical which will no longer add monomer units.



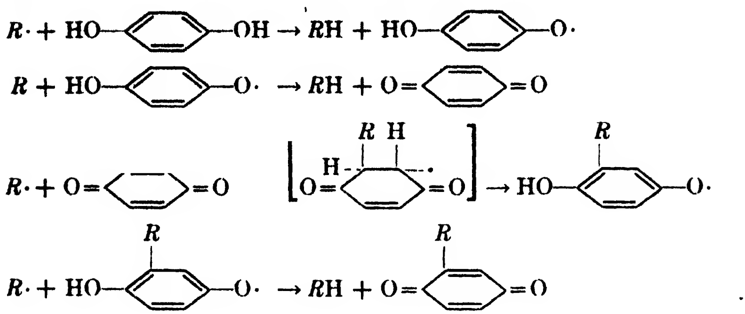
Aromatic nitrocompounds have been found to act as retarders[27] and it has been observed that polystyrene prepared in the presence of nitrobenzene or 2,4-dinitrochlorobenzene contained fragments from the retarding substance in the proper proportion for one fragment per polymer molecule.[11]

$$C_6H_5CH{=}CH_2 \xrightarrow[\substack{(C_6H_5CO_2)_2 \\ 95^\circ}]{(NO_2)_2C_6H_3Cl} C_6H_5CO_2(C_8H_8)_8C_6H_2(NO_2)_2Cl \cdot O$$

Calculated: C, 78.92; H, 6.11; N, 2.39; Cl, 3.03; mol. wt., 1160.
Found: C, 79.02; H, 6.09; N, 2.44; Cl, 2.84; mol. wt., 1180 (visc.).[16]

Substances which act as inhibitors of addition polymerization, such as quinones and hydroquinones, must compete for active radicals much more successfully than substances which act only as retarders.

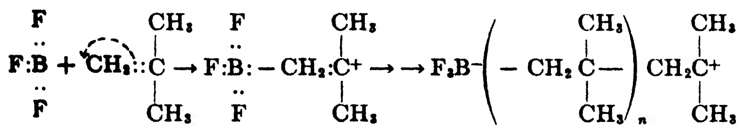
[27] Foord, S. G. Jour. Chem. Soc. 1940: 48



Hydroquinone has been shown to produce quinone in the presence of methyl methacrylate and benzoyl peroxide under condition for polymerization[28] and quinones are known to react with the radicals from acyl peroxides to give alkyl-substituted quinones.[9]

THE CATIONIC MECHANISM

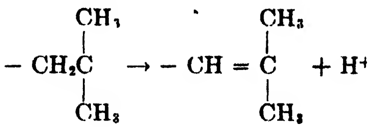
In addition to the many possibilities for free radical type polymerization, there are also many conditions and many monomers for which the reaction undoubtedly proceeds through a mechanism involving an active cationoid intermediate. Thus, such substances as strong acids, boron fluoride, aluminum chloride and stannic chloride, all characterized by strong affinity for a pair of electrons, undoubtedly initiate polymerization by polarization of the double bond of the monomer.[29,30] It is perhaps significant that those substances most readily polymerized by such electrophilic catalysts have substituents such as alkyl, aryl or ether groups, which promote the release of electrons.



The cessation process in this case must be unimolecular rather than bimolecular with respect to active centers. The observed presence of one double bond per polymer molecule for such polymers can be accounted for by loss of a proton from the growing polymer by a process first-order, with respect to the active centers.

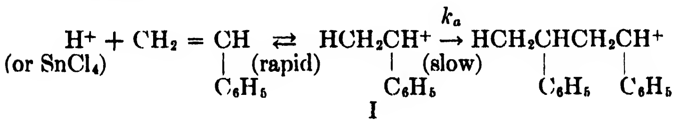
---

[28] Alyea, H. N., Gartland, J. J., & Graham, H. R. Ind. Eng Chem. 34: 458. 1942.
[29] Whitmore, F. C. Ind. Eng. Chem. 26: 94. 1934.
[30] Hunter, W. H., & Yohe, R. V. Jour. Am. Chem. Soc. 55: 1248. 1933.

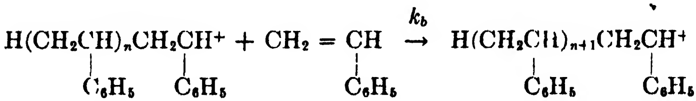


The kinetics of the ionic reaction are markedly different from the free radical type. For example, the polymerization of styrene in thymol has been found to be third-order with respect to monomer.[31] The stannic chloride-catalyzed polymerization[32] was found to be of even slightly higher order with respect to styrene concentration, to be first-order with respect to catalyst concentration, and to produce polymer with degree of polymerization independent of the catalyst concentration. These observations are all satisfactorily accounted for by the following course for the reaction.

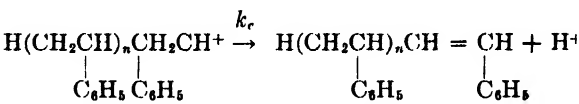
A. Initiation



B Propagation



C. Cessation



If we allow A^+ to represent the active growing chain, the following expressions may be derived:

$$-d[M]/dt = k_b[M][A^+]$$
$$[I] = K_I[M][H^+]$$
$$d[A^+]/dt = k_a[M][I] = k_aK_I[M]^2[H^+]$$
$$-d[A^+]/dt = k_c[A^+]$$

If k_c is very much greater than k_a, a steady state will be reached where the rate of the last two reactions will be equal. The concentration of

[31] Moore, J. G., Burk, R. E., & Lankelma, H. P. Jour. Am. Chem. Soc. 63: 2954. 1941.
[32] Williams, G. Jour. Chem. Soc. 1940: 775.

A^+ may then be expressed in terms of known concentrations of monomer and catalyst and substituted in the equation for the rate of disappearance of monomer.

$$-d[M]/dt = \frac{k_a k_b K_I}{k_c}[M]^3[\mathrm{H}^+]$$

$$\text{chain length} = \frac{k_b[M][A^+]}{k_c[A^+]} = k_b[M]/k_c.$$

RATE THEORY AND SOME PHYSICAL AND CHEMICAL PROPERTIES OF HIGH POLYMERS

By

H. M. Hulburt,* R. A. Harman,† A. V. Tobolsky, and Henry Eyring

From the Frick Chemical Laboratory, Princeton University, Princeton, New Jersey

A. MECHANISM OF POLYMERIZATION

The formation of high polymers from their respective monomers has been the subject of much experimental investigation, much of which has been subject to conflicting interpretations. It is generally conceded that formation of the "addition" type of polymer is a chemical chain reaction. Those molecules which form "addition" polymers are all olefinic or conjugated molecules, and their reactivity is to be associated with the structure of the double bond. Staudinger[1] first proposed that the chain-initiating substance in polymerization is an active monomer molecule in which the double bond has "opened" and can react rapidly. The activation energy for chain initiation would be, on this view, the energy required to form a radical or ion from the monomer. A consideration of the electronic structure of the double bond lends greater precision to our picture of the mechanism by which a double bond "opens." The electronic states of olefins, ethylene in particular, have been considered by Mulliken,[2] Huckel,[3] and Lennard-Jones[4] by the method of molecular orbitals. This theoretical treatment indicates that, whereas ten of the twelve electrons in ethylene are localized by pairs, taking up positions between the carbon and hydrogen nuclei and between the two-carbon nuclei, the remaining two electrons are distributed over the entire molecule. Thus they cannot be said to "belong" to any one carbon atom, for which reason they have been distinguished by the name "unsaturation electrons." Being less tightly bound, these are the electrons which account for the reactivity of the double bond.

The energy states of these electrons are shown in FIGURE 1, where energy is plotted as a function of the angular rotation of one methylene

* National Research Fellow (1942–1943)

† On leave of absence from Allied Chemical and Dye Corporation New York

[1] Staudinger, H. Trans. Faraday Soc 32: 97. 1936, and earlier references cited there

[2] Mulliken, R. S. Rev. Mod. Phys. 14: 265. 1942, and earlier work cited there.

[3] Hückel, E. Zeit. Elektrochem. 43: 752, 827. 1937, and earlier work cited there.

[4] Lennard-Jones, J. E., & Coulson, C. A. Trans. Faraday Soc. 35: 811. 1939, and earlier work cited there.

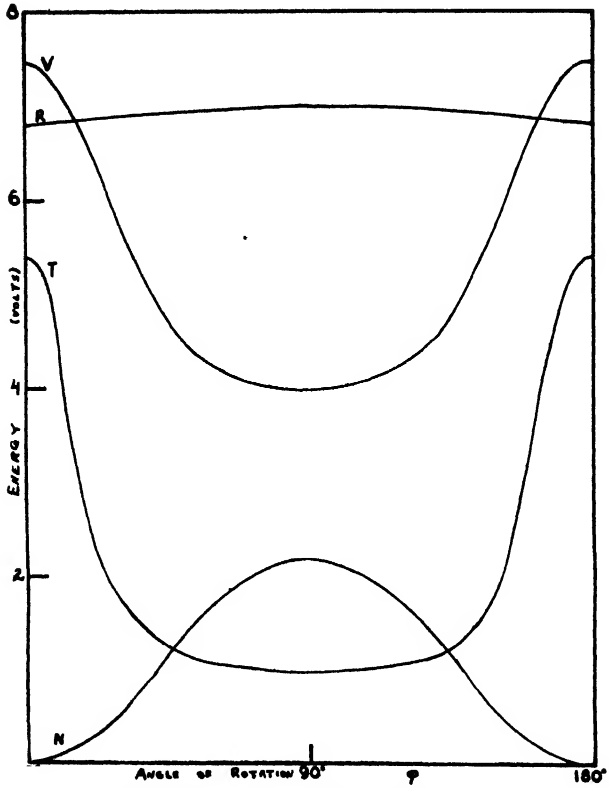


FIGURE 1 Schematic diagram of energy levels in ethylene

with respect to the other In the lowest state, N, the electrons both occupy the same bonding orbital, and hence, by the Pauli Principle, must have antiparallel spins, so that the ethylene is in a singlet state. When one of the electrons is promoted to a nonbonding orbital, two states are possible, a singlet, in which spins are paired, and a triplet, with unpaired spins. In the case of oxygen, which is isoelectronic with ethylene, the triplet state is the ground state, as evidenced by the paramagnetism of oxygen, so that it is not surprising to find the triplet state, T, in ethylene close to the ground state, probably only a volt above it. The corresponding singlet state, V, lies much higher. It is the upper state in the ultraviolet absorption spectrum of olefins which appears in

the region of 1800 Å, the so-called "charge transfer spectrum." The state R is called by Mulliken a "Rydberg" state, since it gives rise to a series of lines in the spectrum similar in appearance to the Rydberg series in hydrogen. In this state, one of the unsaturation electrons is promoted so that it revolves around both carbon atoms much as the electrons in an atom revolve about their nucleus.

The wave functions for these states can be written down in the LCAO molecular orbital approximation. They are:

$$\psi_N = \tfrac{1}{2}(x_A + x_B)_1(x_A + x_B)_2\left(\frac{\alpha_1\beta_2 - \alpha_2\beta_1}{\sqrt{2}}\right), \tag{1}$$

$$\psi_T = \frac{1}{2\sqrt{2}}\left[(x_A + x_B)_1(x_A - x_B)_2 - (x_A + x_B)_2(x_A - x_B)_1\right]\alpha_1\alpha_2, \tag{2}$$

$$\psi_V = \frac{1}{2\sqrt{2}}\left[(x_A + x_B)_1(x_A - x_B)_2 + (x_A + x_B)_2(x_A - x_B)_1\right]\left(\frac{\alpha_1\beta_2 - \alpha_2\beta_1}{\sqrt{2}}\right), \tag{3}$$

where x_A and x_B are $2p_x$ atomic orbitals on the carbon atoms A and B respectively. Multiplying out the space parts of these functions, we see that for $\varphi = 0$ (both methylenes in the same plane):

$$\psi_N = x_{A1}x_{A2} + x_{B1}x_{B2} + x_{A1}x_{B2} + x_{A2}x_{B1}, \tag{4}$$

$$\psi_T = x_{B1}x_{A2} - x_{A1}x_{B2}, \tag{5}$$

$$\psi_V = x_{A1}x_{A2} - x_{B1}x_{B2}. \tag{6}$$

In addition to these states, there is the state in which both unsaturation electrons are in antibonding orbitals:

$$\psi_{N'} = \frac{1}{2}(x_A - x_B)_1(x_A - x_B)_2\left(\frac{\alpha_1\beta_2 - \beta_1\alpha_2}{\sqrt{2}}\right).$$

Optical transitions to this state are forbidden by the symmetry of the molecule. However, the functions x_A and x_B change sign at $\varphi = 90°$, so that as one methylene rotates, the state $\psi_{N'}$ becomes lower and lower in energy and at $\varphi = 180°$ is the ground state. Similarly, the state ψ_N increases in energy upon rotation, so that at $\varphi = 180°$ it plays the same role that $\psi_{N'}$ had at $\varphi = 0°$. At $\varphi = 90°$, the point where ψ_N and $\psi_{N'}$ would have equal energy if it were not for resonance, the wave functions for the upper and lower states are linear combinations of ψ_N and $\psi_{N'}$ or, what amounts to the same thing, of the two states:

$$\psi_N + \psi_{N'} = x_{A1}x_{A2} + x_{B1}x_{B2}, \tag{7a}$$

$$\psi_N - \psi_{N'} = x_{A1}x_{B2} + x_{A2}x_{B1}. \tag{7b}$$

As we shall see the polar state (7a) enters largely into the lowest singlet state at $\varphi = 90°$ and is considerably lower than the polar singlet state V, which is antisymmetric with respect to nuclear interchange.

Thus, the ground state is made up of both polar and homopolar functions, whereas the triplet state is purely homopolar. Hence one is led to speak of transition to state T as "internal radical formation." Transition to state T, however, involves uncoupling the spins of the unsaturation electrons, which can only be done by a magnetic field. The internal field of the molecule is very weak, so that this transition occurs only rarely in an isolated molecule, even though sufficient energy is present. In the presence of atoms, radicals, or molecules containing an odd electron, however, this uncoupling is much more probable and may be expected to occur readily whenever the "odd molecule" comes within the kinetic theory radius of the ethylene.

This picture of the electronic structure of the double bond has been used to account for the two mechanisms of cis-trans isomerization.[5] In TABLE 1 are listed the reactions for which activation energies have been measured. It is clear that they fall into two classes. In the first class are those which have an activation energy of about 42 kcal/mole and a frequency factor of $\sim 10^{11}$, a normal value. In the second class are those isomerizations which have an activation energy of $\sim$23 kcal/

TABLE 1

A. CIS-TRANS ISOMERIZATION

(a) Reactions by the "singlet" mechanism

Compound	Temperature	Pressure	Frequency Factor	Energy
	°C.	mm.	cc./mol. sec.	kcal.
Methyl cinnamate[6] (g)	290–387	5–500	3.5×10^{10}	41.6
-Cyanostyrene[7] (g)	308–378	20–450	4×10^{11}	46.0
Stilbene[8] (g)	280–335	4–400	6×10^{12}	42.8
Stilbene[9] (l)[a]	214–223		2.7×10^{10}	36.7
Monochlorstilbene[9] (l)[b]	226–246		1.4×10^{11}	37.0
Dichlorstilbene[9] (l)	175–196		9.9×10^{10}	34.1

[a] The isomerisation is catalysed by BF_3 at room temperature (reference 16), by HBr and a peroxide (equal Br atoms)[a] (reference 17) and by HBr and a ferromagnetic metal (reference 18).
[b] Catalysed by HBr–oxygen (reference 9).

5 Magee, J. L., Shand, W., & Eyring, H. Jour. Am. Chem Soc. 63: 677. 1941
6 Kistiakowsky, G. B., & Smith, W. R. Jour. Am Chem. Soc 57: 269. 1935.
7 Kistiakowsky, G. B., & Smith, W. R. Jour Am. Chem. Soc. 58: 2428. 1936.
8 Kistiakowsky, G. B., & Smith, W. R. Jour. Am. Chem. Soc 56: 638. 1934.
9 Taylor, T. W. J., & Murray, A. R. Jour. Am Chem Soc. 1938: 2078.
10 [illegible], K. Jour. physikal. Chem. 28: 758. 1924.
11 Kistiakowsky, G. B., & Nelles, M. Zeit. physikal. Chem., Bodenstein Festband 369. 1931.
12 Nelles, M., & Kistiakowsky, G. B. Jour. Am. Chem. Soc. 54: 2208. 1932.

TABLE 1 (*Continued*)

(b) Reactions by the "triplet" mechanism

Compound	Temperature	Pressure	Frequency Factor	Energy
	° *C.*	*mm.*	*cc./mol. sec.*	*kcal.*
Maleic acid[10] (l)	140–150		1.67×10^4	15.8
Dimethyl maleate[9, 11, 12] (g)[a]	270–380	45–4070	6.8×10^5	26.5
[13, 14]	270–320	15–90	3.2×10^3	17.7
Dimethyl citraconate [8] (g)	280–360	30–500	about 7×10^6	25.0
Butene-2 [15] (g)	390–420	100–1440	2	18.0

[a] Catalysed both by Friedel-Crafts catalysts such as $AlCl_3$, $FeCl_3$, $ZnCl_2$ and by paramagnetic substances such as Na^0, O_2, NO (references 13, 14, 19).

B. POLYMERIZATION

(a) Reactions by the "singlet" mechanism

Compound	Temperature	Catalyst	Pressure	Frequency Factor	Energy
	° *C.*			*cc./mol. sec.*	*kcal.*
Ethylene[20] (g)	350–500		2.5–10 atm.	10^{10}	35
Ethylene [21a] (g)	377–393		1420 mm.	4×10^{12}	43.5
Ethylene [21b] (g)	358–414		10 atm.	$10^{13\ 8}$	42.7
Ethylene [21b] ($C_{10}H_8$ soln)	358–414		200–400 atm.	$10^{13\ 4}$	40.6
Propylene[22] (g)	330–400		sealed tube	$10^{11\ 45}$	37.4
Propylene [23]	600–725		atm.	—	41.5
Butene-2 [24] (g)	330–400		sealed tube	$10^{11\ 2}$	38.0
Isobutene [24] (g)	330–400		sealed tube	$10^{13\ 6}$	43.0
Isobutene [25] (l)	—78	BF_3			10
Amylenes[24] (g)	330–400		sealed tube	$10^{11\ 0}$	38.0
Hexene-2[26] (g)	330–400		sealed tube	$10^{10\ 9}$	38.0
Cyclohexene[27] (g)	370–440		sealed tube	$10^{13\ 4}$	47.0
Octene[27] (g)	345–380		sealed tube	$10^{11\ 75}$	40.5
2,3 Dimethyl butadiene-1,3 [28] (g)	309–398		160–195 mm.	1.4×10^{10}	25.3
1,3 Pentadiene [28] (g)	279–419		220–315 mm.	3.5×10^{10}	26.0
Cyclopentadiene [28] (g)	120.4–160		373–1880 mm.	8.5×10^7	14.9
Cyclopentadiene [28] (l)	0–81			2×10^9	17.1
[28] ($C_{10}H_{12}$ soln)	45–71			4.2×10^9	17.3
Chloroprene [29] (l)	40–50			$10^{9\ 45}$	20.2
Styrene [30] (l)	0–38	$SnCl_4$		—	3

13 (a) **Tamamushi, B., & Akiyama, H.** Zeit. Elektrochem. **43**: 156. 1937; (b) **45**: 72. 1939.
14 **Tamamushi, B., & Akiyama, H.** Bull. Chem. Soc. Japan **12**: 382. 1937.
15 **Kistiakowsky, G. B., & Smith, W. R.** Jour. Am. Chem. Soc. **58**: 766. 1936.
16 **Price, C. C., & Meister, M.** Jour. Am. Chem. Soc. **61**: 1595. 1939.
17 **Kharasch, M. S., Mansfield, J. V., & Mayo, F. R.** Jour. Am. Chem. Soc. **59**: 1155. 1937
18 **Urusibara, Y., & Simamura, O.** Bull Chem. Soc. Japan **13**: 566. 1938.
19 **Gilbert, W. I., Turkevich, J., & Wallis, E. S.** Jour. Org. Chem. **3**: 611. 1939.
20 **Pease, R. N.** Jour. Am. Chem. Soc. **53**: 613. 1931.
21 (a) **Storch, H. H.** Jour. Am. Chem. Soc. **57**: 2598. 1935; (b) **Russell, R. P., & Hottel, H. C.** Ind. Eng. Chem. **30**: 183. 1938.
22 **Kranse, M. V., Nemtsov, M. S., & Soskina, E. A.** Jour. Gen. Chem. USSR **5**: 843. 1935.
23 **Moor, W. G., Strigaleva, N. W., & Frost, A. W.** Jour. Gen. Chem. **7**: 860. 1937.
24 **Kranse, M. V., Nemtsov, M. S., & Soskina, E. A.** Jour. Gen. Chem. USSR **8**: 1303. 1938.
25 **Thomas, R. M., Sparks, W. J., & Frolich, P. K.** Jour. Am. Chem. Soc. **62**: 276. 1940.
26 **Nemtsov, M. S., & Polstaev, A. V.** Jour. Gen. Chem. USSR **6**: 892. 1936.
27 **Nemtsov, M. S., Nizovkina, T. V., & Soskina, E. A.** Jour. Gen. Chem. USSR **6**: 1303 1936.
28 **Harkness, J. B., Kistiakowsky, G. B., & Mears, W. H.** Jour. Chem. Phys. **5**: 682. 1937.
29 **Medvedev, S., Chilikina, E., & Klimenkov, V.** Acta Physicochim. USSR **11**: 751. 1939.
30 **Williams, G** Jour. Chem. Soc. **1940**: 775

TABLE 1 (*Continued*)

(b) Reactions by the "triplet" mechanism

Compound	Temperature	Catalyst	Pressure	Frequency Factor	Energy
	° C.			*cc./mol.sec.*	*kcal.*
Styrene [31] (l)	100–132	none		7.5×10^6	28.6
Styrene [32] (l)	27–50	B_2O_2		6×10^{12}	29.3
Styrene [33] (l)	50–60	$(C_6H_5)_3CN_2C_6H_5$ (free radicals)		$10^{12.2}$	23.5
Styrene [33] (l)	90–120			10^7	28.0
d-s-Butyl-chloracrylate [34] (dioxane soln.)	26–28	B_2O_2		$10^{7.7}$	15.2
Indene [35] (l)	120–188			$10^{6.85}$	26.0
Indene [36] (toluene soln.)	120–200			$10^{4.6}$	20.0

C. REACTIONS SHOWING A DUAL MECHANISM DEPENDING ON THE TEMPERATURE

(a) Cis-trans isomerization reactions

Compound	Temperature	Pressure	Frequency Factor	Energy
	° C.		*cc./mol. sec.*	*kcal.*
Dichlorethylene [37] (g) [d]	287–335	200–760 mm.	4.9×10^{12}	41.9
Dichlorethylene [38] (g) [e]	about 200		2.1×10^2	16.0

(b) Polymerization reactions

Compound	Temperature	Pressure	Frequency Factor	Energy
	° C.		(*cc./mol. sec.*)	*kcal.*
Isoprene [39] (g)	268–371	212–739 mm	$2.19 \times 10^{10}\sqrt{T}$	28.9
Isoprene [40] (l)	100–150		$10^{2.8}$	18.8
Isoprene [41] (l)	154–160	1000 atm	$10^{4.84}$	17.0
Butadiene [42] (g)	173–386.4	531–5000 mm.	9.2×10^9	23.7
Butadiene [43] (l or g)	150–200		$10^{9.8}$	25.0
Butadiene [43] (l) [f]	100–150		$10^{3.5}$	13.0

[d] Catalysed by oxygen or air (reference 37).
[e] Catalysed by oxygen and nitric oxide (reference 38).
[f] Catalysed by oxygen (reference 43).

[31] Schulz, G. V., & Husemann, E. Zeit. physikal. Chem. B43: 385. 1939.
[32] (a) Schulz, G. V., Dinglinger, A., & Husemann, E. Zeit. physikal Chem. B39: 246 1938
(b) Schulz, G. V. Naturwiss. 27: 659. 1939.
[33] Foord, S. G. Jour. Chem. Soc. 1940: 48.
[34] Price, C. C., & Kell, R. W. Jour. Am. Chem Soc 63: 2798. 1941.
[35] Breitenbach, J. W. Zeit. Elektrochem. 43: 323 1937.
[36] Dostal, H., & Raff, R. Zeit. physikal. Chem. B32: 417. 1936.
[37] Jones, J. L., & Taylor, R. L. Jour. Am. Chem. Soc 62. 3480. 1940.
[38] Tamamushi, B., Akiyama, H., & Isii, K. Zeit. Electrochem. 47: 340. 1941
[39] Vaughan, W. E. Jour. Am. Chem. Soc 55: 4109. 1933.
[40] Gapon, E. N. Jour. Russ. Phys. Chem. Soc. 62: 1385, 1395. 1930.
[41] Tammann, G., & Pape, E. Zeit. Anorg. Allgem. Chem. 200: 113. 1931.
[42] Kistiakowsky, G. B., & Ransom, W. W. Jour. Chem. Phys. 7: 725. 1939
[43] Lebedev, S. V., Khokhlovkin, M. A., Kulbina, N. I., & Begetova, A. P. Jour. Phys. Chem USSR 7: 130. 1936.

mole and a low frequency factor of $\sim 10^4$. This first class has been supposed to isomerize by rotation against the high energy barrier of the normal state of ethylene, the top of which must consequently be placed at $\sim$42 kcal. The second class of isomerizations is supposed to proceed adiabatically from the state N to the state T at their crossing point. This requires an activation energy of only 23 kcal, but when the two energy surfaces approach very closely, as they do in the uncatalyzed reaction, the probability of remaining in the state N is very high. Thus the transmission coefficient for reaction by passing through the triplet state, which is reflected in the frequency factors observed experimentally, is low.

Examination of a compilation of the rates of initiation reactions in polymerization (TABLE 1) reveals a similar classification. In general, alkyl-substituted olefins polymerize at rates with normal frequency factors and activation energies about 40 kcal/mole which vary with the substituent. This class of polymerizations is catalyzed by generalized acids such as $AlCl_3$, BF_3 and concentrated H_2SO_4, with a marked reduction of the activation energy. The second group comprises styrene and its derivatives and acrylates. All members of this group apparently have the same activation energy for chain initiation. The frequency factor is low for all the uncatalyzed reactions, but is normal ($\sim 10^{11}$) for those catalyzed by free radicals, benzoyl peroxide, and triphenyl methyl diazobenzene. A few substances fall into the first class at one temperature and into the second at lower temperatures. This is natural if there are two competing reaction mechanisms, since the one with the higher activation energy will be favored at high temperatures.

Although alkyl-substituted olefins isomerize by the "triplet" mechanism, they polymerize by the "ionic" mechanism. Similarly, the phenyl-substituted olefins isomerize by the "singlet" mechanism but polymerize by the "triplet" or radical mechanism. This does not really involve a contradiction, however, since those substances which show cis-trans isomerism are necessarily α,β substituted ethylenes which do not polymerize readily. This point of view presupposes that styrene, could its isomerization be detected, would be found to isomerize by the "triplet" mechanism.

Thus there is a large body of fact which can be correlated and understood in relation to the electronic structure of the double bond. We shall consider first the possibility of an "ionic" mechanism in which the growing polymer has a predominantly polar configuration.

Ionic Mechanism

The criteria of an ionic reaction mechanism are generally taken to be:

(1) Marked change in rate upon changing electronegativity of substituents;

(2) Acid or base catalysis;

(3) Intra-molecular rearrangement.

All of these effects are exhibited in polymerizations. The effect of substituents has been discussed often.[44] Asymmetrical substitution of olefins in general enhances the ease with which they polymerize, but substitution of both carbon atoms of the double bond markedly reduces the polymerizability. In dienes, substitution in the 2 or 3 positions enhances polymerizability, whereas substitution in the 1 or 4 positions reduces it greatly. This marked influence of substituents upon reaction rate is easily understood if a polar activated complex is assumed in the reaction. The ease with which such a complex can be formed is related to the degree to which the substituent groups influence the electron charge distribution in the molecule so as to leave one of the carbon atoms of the double bond more positive than the other. Such an atom is then the point of attack for an entering negative group. These concepts are well-developed in the correlation of acid strengths with molecular structure[45] and in the theory of the rates of nitration of substituted benzenes.[46] Polarization of the double bond has been postulated also in the addition of halogens and halogen acids to olefins,[45] so that it is not unreasonable to suppose such a mechanism to operate in addition polymerization also. Thus, the great ease with which vinyl chloride polymerizes as compared with ethylene can be attributed to the polarizing effect of the chlorine on the double bond. This can be expressed by saying that the structures

$$\underset{\text{(induction)}}{\overset{-}{\mathrm{Cl}}-\overset{+}{\mathrm{CH}}=\overset{+}{\mathrm{CH}_2}} \qquad \underset{\text{(resonance)}}{\overset{+}{\mathrm{Cl}}=\overset{+}{\mathrm{CH}}-\overset{-}{\mathrm{CH}_2}}$$

contribute much more to the activated state for this compound than do the corresponding structures for ethylene.

A similar but considerably larger substituent effect is observed in the shift of the spectra of substituted ethylenes.[2] Substitution of methyl for one hydrogen in ethylene decreases the separation between the ground state N and the polar excited state V by 12 kcal/mole.[2] The

[44] See, for example, Burk, R. E., Thompson, H. E., Weith, A. J., & Williams, I. "Polymerization." Reinhold. New York. 1937. Chap. 2.

[45] See Hammett, L. P. "Physical Organic Chemistry" McGraw-Hill. New York. 1940. Magee, J. L., Ri, T., & Eyring, H. Jour. Chem. Phys. 9: 419. 1941.

[46] Ri, T., & Eyring, H. Jour. Chem. Phys. 8: 433. 1940.

activation energy for ionization is reduced only by about 0.5 to 1.0 kcal/mole at ordinary temperatures.[45] Moreover, the excitation energy for the $N \rightarrow V$ transition in the spectrum of plane ethylene is 7.61 e.v., whereas the maximum activation energy observed for the polymerization of ethylene is about 2 e.v. (43,000 cal).[47] This indicates that the polar state involved in thermal polymerizations is not the state, V, which is the upper level in absorption spectra, but rather the state, 7a.

In the absence of detailed studies of those polymerizations where an ionic mechanism is suspected, it is possible to correlate reaction rate with molecular structure only very roughly. However, it is to be noted that among the polymerizations listed in TABLE 1 as "singlet mechanism," there is a wide variability in the rate as well as in the activation energy. There is evidence that the diene polymerizations proceed under certain conditions by a radical mechanism and, as is seen in section B of TABLE 1, the activation energies are low, so that their inclusion under "singlet mechanism" is at best uncertain Those polymerizations catalyzed by peroxides or by free radicals, as well as those which presumably proceed by the "triplet" state in the absence of catalyst, have activation energies which fall within a remarkably narrow range. This observation strengthens the view that the latter proceed by a free radical chain, in which charge distribution would have a very minor effect on activation energy, whereas the former react via a polar activated complex

The second criterion of a polar mechanism, viz., acid or base catalysis, is also widely observed in addition polymerization. Many commercial processes employ catalysts such as $AlCl_3$, BF_3, or concentrated H_2SO_4, all of which may be classed as acids in the sense that they are capable of forming a bond with any molecule which can furnish a pair of electrons. The unsaturation electrons of the olefin or diene are ideally adapted for this purpose, as is evident from the large number of complexes formed between $AlCl_3$ and unsaturated molecules.[48] The fact that many of these are colored indicates the stabilizing effect of the $AlCl_3$ on the ionic states of the molecules involved, since color can arise only from the reduction in the excitation energy of the polar state.

Acidic catalysts exert a much greater effect on the activation energy of polymerization than do peroxides. In the case of styrene, catalysis by benzoyl peroxide leaves the activation energy practically unchanged, but raises the frequency factor. Catalysis by $SnCl_4$ in CCl_4 solution, however, reduces the activation energy to less than three kilocalories[49]

[47] Storch, H. H. Jour. Am. Chem. Soc. 57. 2598. 1935.
[48] Thomas, C. A "Anhydrous $AlCl_3$ in Organic Chemistry" Reinhold. New York. 1941. P. 48.
[49] Williams, G. Jour Chem Soc 1940· 775.

from the uncatalyzed value of ~24 kcal. This indicates a marked difference in the mechanisms of the polymerization in the two cases. Although the polymerization of styrene and vinyl acetate has been found to be uninfluenced by surface when they proceed by the "triplet" mechanism, the polymerization of dienes has been found to be heterogeneous and very sensitive to surface conditions. This fact may also indicate an ionic mechanism in which the silica surface acts as an ionizing two-dimensional solvent, as has been suggested[50] for catalysis of dehydration and hydrogenation on metal oxide surfaces. The polymerization of ethylene has been reported to yield predominately propylene in a silica tube,[51] whereas butylene is reported as the chief product in a pyrex or copper vessel.[52] This suggests that the nature of the ionic surface is important in determining the reaction course, as it would be if ionic-activated states were involved.

The third criterion, molecular rearrangement, is also observed in the dimerization and polymerization of many substances. The experimental evidence for rearrangement before or after polymerization is extensive and will not be reviewed here.[53] The work of Whitmore and associates[54] on the dimers and trimers formed by the action of concentrated sulfuric acid on secondary and tertiary butyl and amyl alcohols furnishes strong evidence for the polar character of these reaction mechanisms.

All these lines of evidence lead one to suspect that addition polymerizations can proceed by a polar mechanism. Closer consideration of the ground state of ethylene indicates that this mechanism is also suggested by the electronic structure. As we have seen, the ground state of plane ethylene is a linear combination of the polar and homopolar states (7a) and (7b). The exact proportions in which they are mixed depends upon the difference in their energies, the state with lowest energy predominating. In plane ethylene, the LCAO molecular orbital approximation gives a ground state, ψ_N, which is 50 per cent polar. As one methylene is rotated with respect to the other, the homopolar C–C bond made by the two unsaturation electrons is weakened until at $\varphi = 90°$ it is practically nonexistent. The activation energy for this rotation, from isomerization experiments, is about 45,000 cal. or 2 e.v.

For ethylene in the perpendicular form, the difference in energy between the polar and the homopolar states can be estimated by consider-

[50] Eyring, H., Hulburt, H. M., & Harman, R. A. Ind. Eng. Chem. 35: 511. 1943.
[51] Frey, F. E., & Smith, D. F. Ind. Eng. Chem. 20: 948. 1928.
[52] Pease, R. N. Jour. Am. Chem. Soc. 53: 613. 1931.
[53] Bergmann, E. Trans. Faraday Soc. 35: 1025. 1939.
[54] Whitmore, F. C., et al. Jour. Am. Chem. Soc. 63: 756, 1120, 1460, 2035, 2197, 2200. 1941.

ing the effect of localizing two *p*-electrons on one carbon atom. This electron transfer reduces the bonding of the electrons to the nucleus by partially shielding it. To balance this, there is an increase in binding energy because of the electrostatic attraction between the nuclei. The first effect can be estimated roughly by the use of the Slater screening constants. Thus, addition of one electron screens the carbon atom by .30, decreasing the electronic binding by $[1 - (.70)^2]I$, where I is the ionization potential of the electron. For plane ethylene, $I = 10.45$ e.v., but this must be reduced by 2 e.v. for the perpendicular form. Thus the total reduction in binding due to transfer of two electrons is $2(1 - .49)I = 8.67$ e.v. The electrostatic potential energy of a positive and a negative charge separated by 1.54 Å is approximately 1/3 that of the hydrogen atom, where the charges are separated by only .53 Å. The increase in total energy due to electrostatic bonding is thus 2/3 the ionization potential of hydrogen or $2/3 \times 13.65 = 9.1$ e.v. Thus there is an apparent gain of 0.43 e.v. binding energy in the polar state.

While no stress can be placed on the exact magnitude of the numbers in this very rough estimate, it is evident that the polar state will contribute largely to the lowest singlet state of ethylene in the perpendicular form. The actual configuration in the activated state, while not completely ionic, since homopolar states undoubtedly contribute in some degree to the resonance structure, is probably predominantly polar in nature. Thus we are led to speak of such an activated complex as "internally ionized."

The substituent effect is clearly evident, since the energy of the ground state (I in the above calculation) rises from -10.45 e.v. to -8.3 e.v. as one to four methyl groups are substituted in the ethylene. Since the electrostatic attraction does not depend markedly upon the substituent, it is evident that increasing substitution increases the polar character of the activated state.

The catalysis by acids is caused by the bonding effect of the acid for the unsaturation electrons, stabilizing them on one carbon atom. This, of course, increases the contribution of polar terms in the ground state, and may be expected to have an even larger effect on the activated state.

In the perpendicular form, the orbitals of the unsaturation electrons of ethylene are directly opposite those bonding the substituents on the carbon atoms. There should consequently be some overlap between them, with the result that the substituents may be considered to be partially bound to both carbon atoms of the double bond. Such a *three-center two-electron bond* is similar to that of H^+_3, where it contributes

great stability to the molecule.[55] In the case of ethylene, the three centers, carbon, carbon and substituent, share the two unsaturation electrons with a consequent partial release of the usual substituent-bonding electrons to give an increased electron density on one carbon atom. This state might be represented thus:

$$\begin{array}{ccccc} & & & \mathrm{H} & & \mathrm{H} \\ {}^{-}\mathrm{C} & & : & & \grave{\mathrm{C}}_{+} \\ \ddot{\mathrm{H}} & & & & \ddot{\mathrm{H}} \end{array}$$

The substituents on the negative carbon atom tend to migrate to the positive carbon. Such three-center bonds have been proposed in the mechanism of bromine addition to olefins[56] as well as in the addition of ions to double bonds[57] in the formation of olefin-ion complexes in solution. This type of bonding affords a more detailed mechanism for the intramolecular rearrangement studied extensively by Whitmore and co-workers.[55]

Whitmore has proposed that acid-catalyzed polymerizations have an ionic mechanism.[58] This viewpoint can now be extended and elaborated. We have seen that, even in the absence of catalyst, the activated state of the double bond is probably highly polar in structure. Thus two ethylene molecules may be supposed to react to form an internally ionized dimer.

$$H_2C{:}\,{:}CH_2 + H_2C{:}\,{:}CH_2 - {}^{-}{:}\overset{\mathrm{H}}{\underset{\mathrm{H}}{\mathrm{C}}}{:}\overset{\mathrm{H}}{\underset{\mathrm{H}}{\mathrm{C}}}{:}\overset{\mathrm{H}}{\underset{\mathrm{H}}{\mathrm{C}}}{:}\overset{\mathrm{H}}{\underset{\mathrm{H}}{\mathrm{C}}}{}^{+}$$

In this dimer, the substituents on the carbon adjacent to the positive charge tend to become bonded to the positive carbon, so that migration is relatively easy Thus, the dimer may become stabilized by substituents shifting by one carbon atom down the chain until the double bond is regenerated at the other end. Competing with this unimolecular process is the addition of the internally ionized dimer to another molecule of monomer to give ionic trimer, leading the growth of long chains. Growth can proceed with little activation energy so long as the double bond is not regenerated, since the polymer is already in the polar activated state.

...felder, J. O. Jour. Chem. Phys. 6: 795. 1938.
...rts, I. R., & Kimball, G. E. Jour. Am. Chem. Soc. 59: 947. 1937.
...s, H. J., Moore, R. S., & Pressman, A. Jour. Am Chem Soc 65: 227. 1943. See also
...s, M. G. Trans Faraday Soc. 38: 340, 348, 356. 1942.
...more, F. C. Ind. Eng. Chem. 26: 94. 1934

These reactions present all the essential requirements for a chemical chain leading to long polymer molecules. The fact that many substances form only short chains is to be attributed to the speed with which rearrangement occurs in competition with addition. Since rearrangement is unimolecular, whereas addition is bimolecular, long chains will be favored in condensed phases. Long chains will be favored at low temperatures provided the activation energy for termination is greater than that for growth of the chain. This seems to be the case in the $AlCl_3$-catalyzed polymerization of isobutene, which forms the longest chains at the lowest temperatures.

These reactions lead to the following kinetic scheme, in which A represents monomer and B the catalyst:

$$A + B \xrightarrow{k_1} (AB)^*$$

$$(AB)^* + A \xrightarrow{k_2} (BA_2)^*$$

etc.

$$(BA_n)^* \xrightarrow{k_3} A_n + B.$$

This reaction mechanism gives the over-all rate

$$-\frac{dA}{dt} = k_2 \frac{k_1}{k_3} A^2 B$$

and the degree of polymerization

$$P = \frac{k_2 A}{k_3}$$

In the absence of catalyst, B should be replaced by A in these equations.

The published kinetic data on polymerization by ionic mechanism are sparse and not entirely consistent. The polymerization of styrene in carbon tetrachloride solution has been studied in the presence of stannic chloride.[59] The reaction shows an induction period, and is poisoned by HCl, and by styrene dimer. The specific viscosity of the resulting polymer is proportional to the square root of the concentration of styrene, but is independent of the catalyst concentration. The maximum rate is proportional to the first power of the catalyst and to the fourth power of the monomer. These results are difficult to harmonize with the proposed mechanism. They would require a third-order initiation, which seems highly improbable. However, work of Moore, Burk and Lankelma[60] on the polymerization of styrene in thymol

[59] Williams, G. Jour. Chem. Soc 1938: 247, 1048. 1940: 775.
[60] Moore, J. E., Burk, R. E., & Lankelma, H. P. Jour. Am. Chem. Soc. 63: 2954. 1941

solution gives data in good agreement with these equations. There was no evidence that thymol was incorporated in the polymer. Its action may be interpreted rather as that of a polar solvent which stabilizes the internal ionization of the double bond by complex formation, so that the ionic mechanism is favored over the "triplet" or radical mechanism which apparently occurs, for example, in toluene solution. The experiments reported by Frohlich and co-workers[61] on the low temperature polymerization of isobutene in the presence of boron trifluoride do not seem to contradict the hypothesis of an ionic mechanism.

Thus it is evident that, while there is abundant qualitative evidence for the possibility of an ionic mechanism in polymerization, there is a great lack of complete kinetic investigations of the polymerization process in the presence of acid catalysts.

Polymerization of Vinyl Compounds

The polymerization of styrene is the most thoroughly studied reaction for which there are published data. The extensive experiments of Schulz[62] and co-workers, and Suess[63] and co-workers present a number of outstanding facts which must be reconciled in any proposed mechanism. In addition, the experiments of Medvedev and Kamenskaya,[64] and Cuthbertson, Gee and Rideal[65] on the polymerization of vinyl acetate show a similar general behavior. We shall discuss first the data on the polymerization of styrene.

The thermal polymerization of styrene in the absence of any intentional catalyst has been extensively studied, both in solution and in the pure liquid.[62, 63] The rate of disappearance of styrene monomer in solution is found to be well represented in seven different solvents studied in two laboratories by the expression

$$-\frac{dA}{dt} = k_0 A^2, \tag{8}$$

where A represents the monomer concentration. In each case, a plot of the logarithm of the initial rate of reaction versus the logarithm of the initial concentration of styrene gives a straight line with a slope of two. The rate of disappearance is not markedly different in different solvents.

[61] Thomas, R. M., Sparks, W. T., Frohlich, P. K., Otto, H., & Mueller-Cunradi, M. Jour. Am. Chem. Soc. 62: 276. 1940

[62] (a) Schulz, G. V., & Husemann, E. Zeit physikal. Chem B43: 385 1939, (b) Schulz, G. V., Dinglinger, A., & Husemann, E. Zeit physikal. Chem. B39: 246. 1938. (c) Schulz, G. V. Zeit. Elektrochem 47: 265 1941, (d) Schulz, G. V, & Husemann, E Zeit physikal Chem B34: 187. 1936.

[63] (a) Suess, H., Pilch, K., & Rudorfer, H Zeit physikal Chem. A179: 361. 1937, (b) Suess, H., & Springer, A. Zeit. physikal. Chem. A181: 81. 1937

[64] Kamenskaya, S., & Medvedev, S. Acta Physicochim. USSR 13: 565. 1940.

[65] Cuthbertson, A. C., Gee, G., & Rideal, E. Proc. Roy. Soc A170: 300 1939.

However, the mean molecular weight, or the degree of polymerization,* of the product varies widely. For example, a solution of monomer in cyclohexane gave a product of degree of polymerization 1105, whereas in ethylbenzene, the same initial concentration of monomer gave a product of d.p. 455. The over-all rates varied only from 1.8×10^{-7} to 2.7×10^{-7} for the same solvents.

When benzoyl peroxide is added to styrene in solution, polymerization is much more rapid and the degree of polymerization is much less. The initial rate of polymerization in toluene, the only solvent studied, is well represented by

$$-\frac{dA}{dt} = k_0 B^{1/2} A^{3/2}, \tag{9}$$

where B is the concentration of benzoyl peroxide initially present. The degree of polymerization is proportional to the square root of the benzoyl peroxide concentration.

These findings are in partial agreement with those for other compounds. The polymerization of vinyl acetate in the presence of benzoyl peroxide has been said[64] to follow a kinetics which would lead to an over-all rate second-order in monomer. However, plotting the logarithm of initial rate versus the logarithm of initial concentration of monomer gives a line of slope no greater than $3/2$. A wider range of concentrations would be desirable to test this point further, but the present data do not indicate a second-order over-all rate. This is confirmed by the work of Cuthbertson, Gee and Rideal,[65] who were led to adopt an equation of the form of equation (9) to express their results. The benzoyl peroxide catalyzed polymerization of d-s-butyl-α-chloracrylate[66] is reported to be first-order in monomer and half-order in benzoyl peroxide. In this case it is to be noted that the catalyst concentration is from 0.1 to 0.5 moles/liter, whereas in the other cases reported above, the benzoyl peroxide concentration is of the order of .05 moles/liter.

The thermal polymerization of pure liquid styrene[62d] has been reported to follow a first-order law, but a recalculation of the data on the basis of a second-order rate gives uniform rate constants out to 75 per cent conversion, whereas the first-order law holds only to 25 per cent conversion. A three-halves-order law holds only up to 50 per cent conversion. It is not inconsistent with the data, therefore, to suppose this case to follow the rate law.

$$-\frac{dA}{dt} = k_0 A^2.$$

* As used in this paper, *degree of polymerization* (d. p. or P) signifies the number of repeating units in the polymer molecule.

[64] Price, C. C., & Kell, R. W. Jour. Am. Chem. Soc. 63: 2798. 1941.

Additional information can be gained by combining data on over-all rates with the data on the degree of polymerization if some assumptions are made as to the mechanism of the reaction chain. The propagation reaction offers no difficulty. Since long chains are formed very rapidly, the velocity of propagation must be rapid compared to that of termination. This will be the case if propagation consists of the addition of monomer to a polymer radical to give a polymer radical one unit longer. The velocity of propagation is then

$$V_{\text{propagation}} = k_2AC^*, \tag{10}$$

where C^* is the total concentration of polymer radicals. For the present case, in which growth is much faster than initiation, the over-all rate of polymerization is just the rate of propagation, i.e.,

$$-\frac{dA}{dt} = k_2AC^*\left[1 + \frac{V_{\text{init}}}{V_{\text{prop}}}\right], \tag{11}$$

where the second term in the bracket is very small.

In the steady state, the velocity of chain initiation must equal that of chain termination. This relationship permits the calculation of the radical concentration, C^*. Moreover, we see that the second term in the bracket in equation (11) may then be written

$$\frac{V_{\text{init}}}{V_{\text{prop}}} = \frac{V_{\text{term}}}{V_{\text{prop}}} = \frac{1}{\gamma}, \tag{12}$$

where γ is the kinetic chain length of the reaction. As will become evident, only rarely is the kinetic chain length, γ, equal to the degree of polymerization, P, although this equality has commonly been assumed in the past. However, if γ is known, combining equations (12), (11) and (10) shows that

$$-\frac{1}{\gamma}\frac{dA}{dt} = V_{\text{init}}. \tag{13}$$

Several possibilities present themselves for the chain termination. Granted a radical chain mechanism, it is hard to visualize any mechanism for termination which does not involve the reaction of two free radicals with each other. We have seen that termination by unimolecular rearrangement or by ring closure probably involves polar intermediates, rather than homopolar free radicals. Reaction of a radical with monomer, or with solvent, only generates a new radical, which can initiate a new polymer growth but does not terminate the kinetic reaction chain. Two polymer radicals, however, may simply unite at their free valences to form a single polymer molecule, thus destroying two radicals. Alter-

natively, one radical may remove a hydrogen atom from the β carbon of the other, leaving two polymer molecules, one saturated and the other unsaturated. In either case, we have

$$V_{\text{term}} = k_3 C^{*2}. \tag{14}$$

In the steady state, we have

$$V_{\text{init}} = V_{\text{term}},$$

whence, by equation (14),

$$C^* = (V_{\text{init}}/k_3)^{1/2}. \tag{15}$$

Substituting in equation (11) gives us

$$-\frac{dA}{dt} = k_2(V_{\text{init}}/k_3)^{1/2}A[1 + (1/\gamma)]. \tag{16}$$

Thus, having fixed the mechanisms of propagation and termination by *a priori* considerations, the mechanism of the initiation step must conform with the observed order of the over-all reaction. The second-order over-all rate observed in the uncatalyzed polymerization of styrene both in solvents and as pure liquid requires second-order initiation. The three-halves-order rate observed in the catalyzed polymerization requires an initiation which is first-order in monomer. The half-order dependence on catalyst requires the initiation to be first-order in benzoyl peroxide. The unimolecular over-all rate, observed by Price,[66] requires initiation independent of monomer concentration. The three cases can all be formulated formally in the expression

$$V_{\text{init}} = k_1 AB, \tag{17}$$

in which we take B equal to the benzoyl peroxide concentration in the catalyzed case and equal to A, the concentration of monomer, in the uncatalyzed case. For the case of unimolecular over-all rate, we must take $A = 1$ and B equal to the catalyst concentration. We have, then, from equations (12) and (16):

$$\frac{1}{\gamma} = \frac{(k_1 k_3 B)^{1/2}}{k_3 A^{1/2}}; \qquad C^* = \left(\frac{k_1 AB}{k_3}\right)^{1/2}; \tag{18}$$

$$-\frac{dA}{dt} = k_2\left(\frac{k_1 B}{k_3}\right)^{1/2} A^{3/2}\left(1 + \frac{1}{\gamma}\right) \tag{19}$$

Degree of Polymerization and Kinetic Chain Length

It has already been remarked that the degree of polymerization is not generally equal to the kinetic chain length, as is often supposed. If this

were the case, we see from equation (18) that the degree of polymerization would be proportional to the square root of the monomer concentration when catalyst is present. The data of Schulz and Husemann[62a] on the catalyzed polymerization of styrene in toluene solution permit a test of this relation. In FIGURE 2 is plotted the logarithm of P (degree of polymerization) versus the logarithm of $A_0^{1/2}$ (initial concentration of monomer) for the initial stages of reaction (less than 5 per cent polymerized). The points deviate slightly from a straight line. Moreover, the best straight line through them has a slope definitely less than unity. This indicates that some reaction is going on which terminates the polymer molecules without terminating the kinetic reaction chain.

The data for uncatalyzed thermal polymerization of styrene are even more unequivocal. In this case, we have $B \equiv A$ in equation (18), and γ, the kinetic chain length, is independent of the monomer concentration. However, the ample data of Schulz, Dinglinger and Husemann[62b]; Suess, Pilch and Rudorfer[63a]; and of Suess and Springer[63b] demonstrate

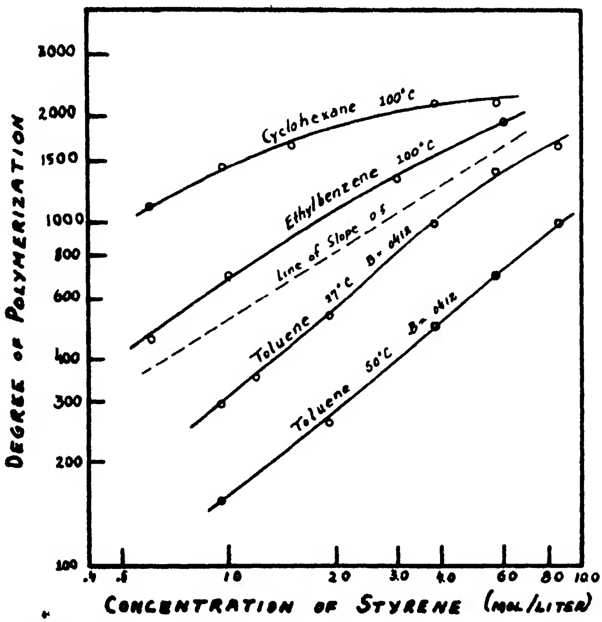


FIGURE 2. Degree of polymerization of styrene as a function of monomer concentration in different solvents. B denotes concentration of benzoyl peroxide

that the degree of polymerization is a function not only of the concentration of monomer but also of the solvent used. This again indicates that polymer growth is terminated without terminating the reaction chain. Medvedev and Kamenskaya[64] reach the same conclusion from their study of vinyl acetate polymerization.

Such reactions were proposed by Flory[67] and termed "chain transfer" reactions. The growing polymer radical supposed to react with a monomer molecule, or with the solvent, so that its growth is terminated and a new radical generated. These reactions may be formulated thus:

$$A_m^* + A \xrightarrow{k_t} A_m + A^*$$

$$A_m^* + S \xrightarrow{k_t'} A_m + S^*$$

$$S^* + A \xrightarrow{k_s} A^* + S.$$

The degree of polymerization of the polymer formed during any (infinitesimally) small amount of reaction is then given by the ratio of the rate at which the reaction chain is propagated to the rate at which growing polymer radicals are terminated by all processes. Thus,

$$P = \frac{V_{\text{prop}}}{V_{\text{term}} + V_{\text{transfer}}} = \frac{k_2 A C^*}{k_3 C^{*2} + k_t A C^* + k_t' S c^*} \tag{20}$$

Substituting from equation (18) and inverting, this becomes

$$\frac{1}{P} = \frac{k_t' S}{(k_2 + k_t)A} + \frac{(k_1 k_3 B)^{1/2}}{(k_2 + k_t)A^{1/2}} + \frac{k_t}{k_2 + k_t}, \tag{21}$$

where S is the solvent concentration. This formula holds only for the polymer formed when the monomer concentration is constant, i.e., at the beginning of the reaction. FIGURE 3 shows that a graph of $1/P$ versus $1/A_0^{\frac{1}{2}}$ passes through the origin for the catalyzed polymerization of styrene dissolved in toluene. Hence k_t must be zero. The only important transfer process is reaction with the solvent. In this case, we have

$$\frac{A^{1/2}}{P} = \frac{k_t' S}{k_2 A} + \frac{(k_1 k_3 B)^{1/2}}{k_2} \tag{22}$$

FIGURE 4 indicates that this relation is obeyed by styrene dissolved in toluene. At small values of $A^{\frac{1}{2}}$, the solvent concentration, S, is no longer approximately unchanging, so that some departure from the initial slope is to be expected in the more concentrated solutions. Since

[67] Flory, P. J. Jour. Am. Chem. Soc. 59: 241. 1937.

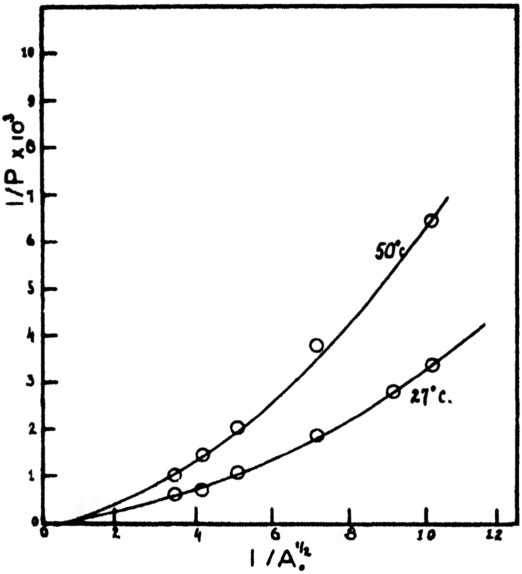


FIGURE 5 Effect of chain transfer on degree of polymerisation of styrene in toluene

the densities of the solutions are not reported, it is impossible to correct for this factor.

In the absence of catalyst, equation (21) becomes (setting $B = A$)

$$\frac{1}{P} = \frac{k_t'S}{k_2A} + \frac{(k_1k_3)^{1/2}}{k_2} \tag{23}$$

if we assume $k_t = 0$, i.e., that transfer to monomer is negligible in comparison with transfer to solvent. FIGURES 5 and 6 show that this relationship is obeyed in many different solvents. Suess and co-workers[62] have showed independently that it is obeyed by the polymerization of styrene in benzene, ethyl benzene, diethylbenzene, heptane, toluene and carbon tetrachloride. In the last solvent, direct chemical evidence that carbon tetrachloride fragments enter into the polymer has been found.[63]

It should be noted that the intercept on the $1/P$ axis, which gives the reciprocal kinetic chain length, $1/\gamma$, is essentially independent of the solvent used. The uncertainty in the extrapolation of the results in the most dilute solution makes it impossible to determine how rigorously

[63] Breitenbach, J. W., & Maschin, A. Zeit. physikal. Chem. A187: 175. 1940.

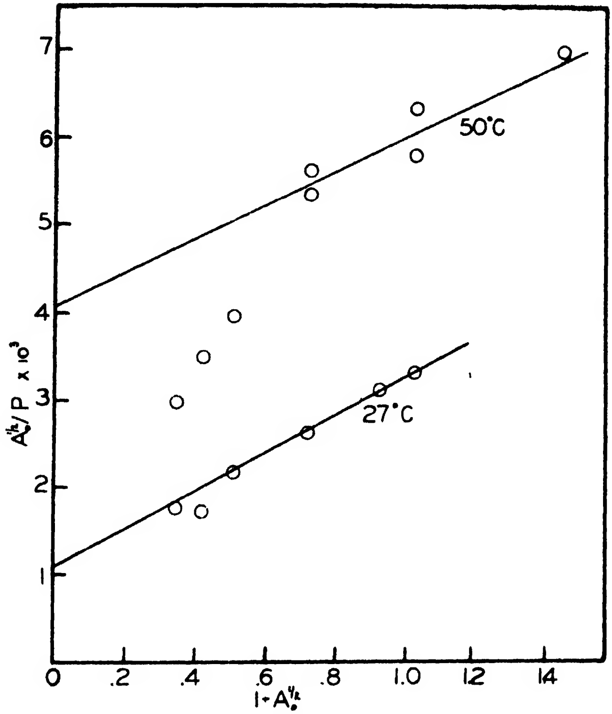


FIGURE 4. Calculation of kinetic chain length in styrene polymerization in toluene

this independence of solvent holds, but it appears highly probable that the solvent has no specific chemical action in chain initiation or in chain propagation except as it may alter the thermodynamic activities of the initial and activated states in these steps. The data of Suess and co-workers is less complete at low concentrations. Consequently the extrapolation is still more uncertain and shows somewhat greater apparent variation in kinetic chain length with solvent. However, there are no differences greater than the large "experimental error" in determining the chain length.

Having determined the kinetic chain length, we may use equation (13) to calculate the velocity of the initiation reaction. The results of these calculations are recorded in TABLE 2 for the cases studied by Schulz and co-workers. All the derived values have been calculated by us from the

TABLE 2

A. POLYMERIZATION OF STYRENE IN TOLUENE, CATALYZED BY BENZOYL PEROXIDE*

$$-\frac{dA}{dt} = k_0 B^{1/2} A^{3/2} \qquad B = .0412 \text{ moles liter}^{-1}$$

	27° C.	50° C.	A	E
k_0	4.953×10^{-7}	6.426×10^{-6}	3×10^{8}	20,300
$\left(\frac{k_1 k_3}{k_2}\right)^{1/2}$	5.420×10^{-3}	2.020×10^{-2}	—	19,010
k_1	2.695×10^{-9}	1.366×10^{-7}	6.0×10^{12}	29,310
$\frac{k_t S}{k_2}$	2.11×10^{-3}	1.9×10^{-3}	—	—

B. UNCATALYZED POLYMERIZATION OF STYRENE IN SOLUTION

Solvent	k_0: $A \times 10^{-6}$	k_0: E	k_1: $A \times 10^{-5}$	k_1: E	$k_t S\ k_2 \times 10^3$: 100° C.	$k_t S\ k_2 \times 10^3$: 132° C.	E
Benzene					.40	1.0	8,860
Toluene	9.15	24,700	7.5	28,630	.52	1.13	7,430
Cyclohexane					.335	.55	4,720
Ethyl benzene							
Diethyl benzene	.11	24,700	9.02	28,630	1.18 2.03	1.57 2.65	2,710 2,620
		100° C	132° C.	A	E		
$\left(\frac{k_1 k_3}{k_2}\right)^{1/2}$		4×10^{-4}	6×10^{-4}	8.2×10^{-2}	3,930		

* All rate constants are in units of cc-mole sec.$^{-1}$

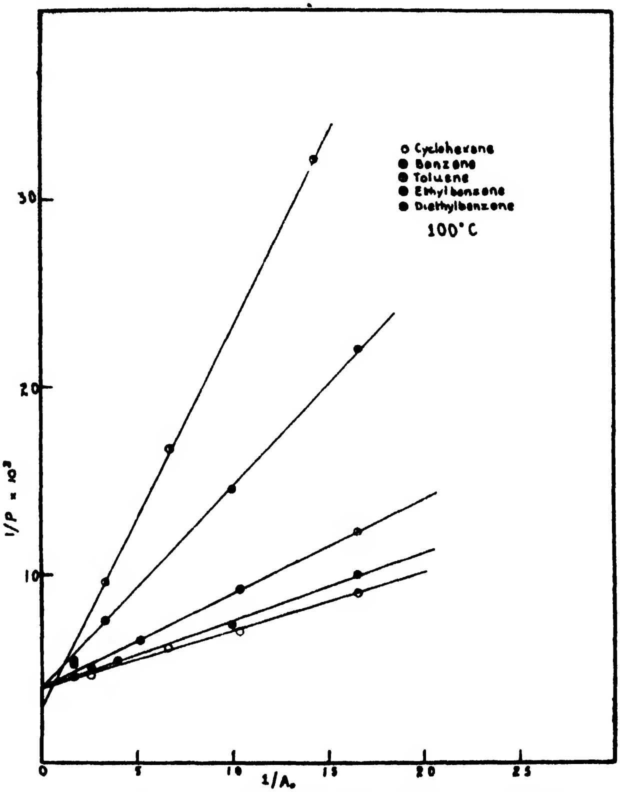


FIGURE 5 Chain transfer in thermal polymerisation of styrene in solvents

initial rates reported by Schulz. It is to be noted that the interpretation given the effect of solvents differs considerably from that proposed by Schulz, who ignored chain transfer completely. For this reason, the values for the activation energy of initiation given in TABLE 2 are higher than those reported by Schulz, and are different in the catalyzed and uncatalyzed cases, contrary to his report. The result for the uncatalyzed reaction is in good agreement with the value of 28 kcal. determined from the rate of disappearance of benzoquinone inhibitor.[89]

The velocity of chain transfer to solvent is seen to be about one one-thousandth that of chain propagation. This would require a difference

[89] Foord, S. G. Jour Chem Soc 1940: 48

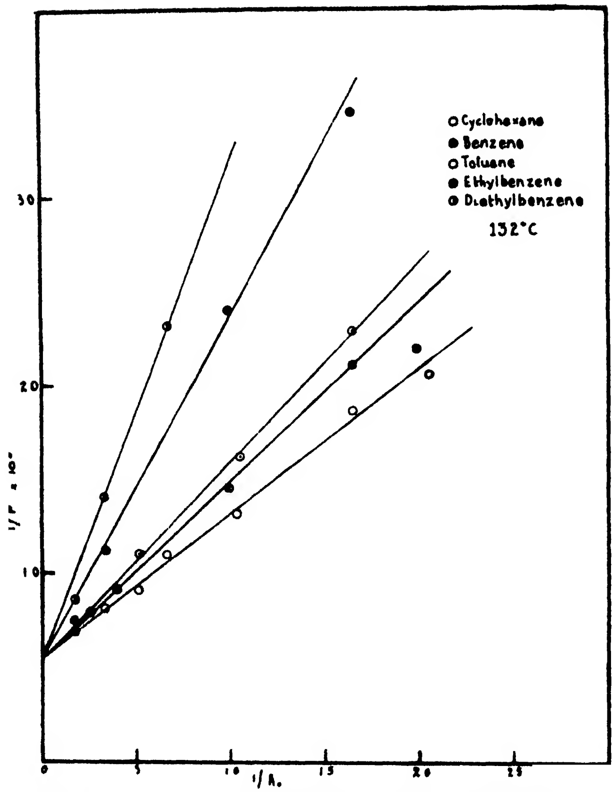


FIGURE 6. Chain transfer in thermal polymerization of styrene in solvents

in activation energies of about 5.5 kcal./mole if the frequency factors of these rate constants were the same. The activation energies calculated from the temperature dependence of $k_t'S'/k_2$ evaluated graphically are of this order of magnitude.

The rate constants for termination and propagation, k_3 and k_2 respectively, occur only in the ratio

$$\frac{k_2}{k_3^{1/2}} = \frac{k_0}{k_1^{1/2}} = 16e^{-4500/RT}.$$

Thus, we have

$$E_2 - \tfrac{1}{2}E_3 = 4500.$$

We see that the activation energy for propagation must exceed 4.5 kcal./mole. The kinetics give us no decision as to whether termination is by recombination of radicals or by disproportionation between them.

The values for activation energy of initiation given in TABLE 2, together with the general theory of the double bond, permit some speculation concerning the mechanism of chain initiation. In the first place, it is apparent that the catalysis by benzoyl peroxide does not consist in a marked lowering of the activation energy for chain initiation, but rather in a raising of the frequency factor. The uncertainties in the activation energies reported in TABLE 2 are sufficiently large that no significance can be ascribed to the difference between the energy for the catalyzed and uncatalyzed initiations. The requirements of experiment can be met in the uncatalyzed case if it is postulated that the energy-consuming step in the activation is the transition from the singlet to the triplet state in styrene. This occurs, however, very infrequently in the absence of external magnetic fields, so that the transmission coefficient for this process is very low. The initiation can be formulated as follows:

$$A \underset{k_a'}{\overset{k_a}{\rightleftarrows}} A_t$$

$$A_t + A \xrightarrow{k_b} A_2^*.$$

Thus, the triplet styrene is not considered to be as reactive as a real free radical, which is formed upon reaction of triplet styrene with monomer. This mechanism leads to the following expression for the velocity of initiation:

$$V_{\text{init}} = \frac{k_b k_a A^2}{k_a' + k_b A}. \tag{24}$$

Initiation will appear to be second-order, as required by the over-all kinetics, provided $k_a' >> k_b A$. This will be the case if k_b has an activation energy three or four kilocalories greater than k_a. This leads to an activation energy of $\sim$24 kilocalories for k_a, which agrees well with that for the "triplet" mechanism in cis-trans isomerization.

The chain initiation in the catalyzed case is open to two interpretations. One might suppose the benzoyl peroxide to form a molecular complex with the styrene, in which it exerts sufficient interaction to facilitate transition to the triplet state. This "triplet" complex then decomposes to form free radicals containing fragments of benzoyl peroxide. This may be formulated:

$$A + B \underset{k_a'}{\overset{k_a}{\rightleftarrows}} (AB)_t$$

$$(AB)_t \xrightarrow{k_b} A_2^*.$$

The velocity of initiation is given by

$$V_{\text{init}} = \frac{k_b k_a AB}{k_a' + k_b} \tag{25}$$

Initiation will be first-order in monomer, as required by the over-all kinetics. The slightly higher activation energy for initiation in the catalyzed case than in the uncatalyzed case may be attributed to the energy required to dissociate the peroxide bond as the $(AB)_t$ complex decomposes, which would increase the activation energy of k_b. Thus the essential function of the catalyst is to increase the equilibrium number of "triplet" styrene molecules by forming a reactive complex with the double bond.

An alternative explanation of the catalysis would suppose that the true chain-initiating radicals are formed by the decomposition of benzoyl peroxide. Under conditions such that most of these phenyl and benzoyl radicals react with solvent or recombine so as not to lead to polymerization, the initiation will appear to be first-order in monomer as well as in catalyst. Thus:

$$B \xrightarrow{k_a} R\cdot + R\cdot'$$

$$R\cdot \xrightarrow{k_a'} \text{inert products}$$

$$R\cdot + A \xrightarrow{k_b} RA^*$$

$$V_{\text{init}} = \frac{k_b k_a AB}{k_a' + k_b A} \tag{26}$$

Again we must suppose k_a' fast with respect to k_b. However, this explanation leads to the difficulty that the activation energy for initiation should be greater than that for the decomposition of benzoyl peroxide into radicals. This decomposition has been reported[70] to have an activation energy of 31,000 calories, which is nearly the upper limit for the chain initiation in styrene. Medvedev and Kamenskaya[64] report an activation energy of 30,000 calories for this same decomposition. It thus appears difficult to reconcile the quantitative aspects of this mechanism. However, the activation energy for initiation in styrene is not sufficiently certain to eliminate this possibility definitely.

[70] McClure, J. H., Robertson, R. E., & Cuthbertson, A. C. Can. Jour. Res. B20: 103. 1942.

The latter explanation is adequate to explain the first-order over-all rate observed in the polymerization of S-butyl-chlorocrylate if we suppose each catalyst fragment active in initiating a chain.†

A fourth possibility must not be overlooked. The abnormally low frequency factor could arise from the presence of an unknown catalyst in concentrations as low as 10^{-6} moles per liter. Burk and Thompson[71] have demonstrated the difficulty of removing all the dissolved oxygen from styrene. In view of the known accelerating effect of oxygen, the low frequency factor might be attributed to traces of oxygen. However, the reproducibility of the rates in at least two different laboratories tends to weaken this contention. Although this possibility cannot be excluded, the present authors believe it unlikely in view of the concordant results of the hypothesis of a "triplet" state in the initiation process.

It is evident that it would be highly desirable from the theoretical standpoint to study polymerization more completely in dilute solutions, since the effect of chain transfer is most easily accounted for under these conditions. However, the existing data seems to be correlated fairly satisfactorily by the "triplet" mechanism here postulated

Molecular Weight Distribution

It is well known that addition polymers are not pure chemical species, but a mixture of homologous hydrocarbons of all chain lengths. This is qualitatively accounted for by the chain mechanism, since there is a finite chance of termination at every step in the addition process. The explicit mechanism proposed here leads to a molecular weight distribution function which can be checked against experiment. This distribution function takes explicit account of chain transfer as well as of the change in chain length as the amount of monomer is depleted. The derivation is based on the following general set of reactions.

$$A + B \xrightarrow{k_1} A^* \tag{27a}$$

$$A^* + A \xrightarrow{k_2} A_2^* \tag{27b}$$

$$A_m^* + A \xrightarrow{k_2} A^*_{m+1} \tag{27c}$$

$$A_m^* + A \xrightarrow{k_t} A_m + A^* \tag{27d}$$

† This reaction showed a low frequency factor ($10^{7.7}$). This suggests that triplet mechanism is operative (for example in the step $(AB)_t \to A^*$ of equation 25), and that the reaction order is really higher than first. Since the latter possibility is not supported by anything in the published data, the low frequency factor in this case must remain unexplained.

[71] Thompson, H. E., & Burk, R. E. Jour. Am. Chem. Soc. 57: 711 1935

$$A_m^* + S \xrightarrow{k_t'} A_m + S^* \tag{27e}$$

$$S^* + A \xrightarrow{k_s} A^* + S \tag{27f}$$

$$A_m^* + A_n^* \xrightarrow{k_D} A_m + A_n \tag{27g}$$

$$A_m^* + A_n^* \xrightarrow{k_C} A_{m+n}. \tag{27h}$$

The molecular weight distribution can be found by integrating the expression for $\frac{dA_m}{dt}$ derived from these equations by the usual steady-state treatment. Thus,

$$\frac{dA_m}{dt} = k_D A_m^* C^* + k_t A_m^* A + k_t' A_m^* S + \tfrac{1}{2} k_C \sum_{j=1}^{m-1} A_j^* A^*_{m-j}, \tag{28}$$

$$-\frac{dA}{dt} = k_2 AC^* \left[1 + \frac{k_1 AB}{k_2 AC^*} + \frac{k_t AC^*}{k_2 AC^*} - \frac{k_t AA^*}{k_2 AC^*} + \frac{k_s AS^*}{k_2 AC^*} - \frac{k_D A^* C^*}{k_2 AC^*} \right], \tag{29}$$

where C^* is the total concentration of all radicals:

$$C^* = \sum_{m=1}^{\infty} A_m^*. \tag{30}$$

Assuming a steady state gives the following conditions:

$$\frac{dA^*}{dt} = 0 = k_1 AB - k_2 A^* A + k_t AC^* - k_t AA^* - k_t' A^* S + k_s S^* A - k_D A^* C^* - k_C A^* C^*, \tag{31}$$

$$\frac{dA_m^*}{dt} = 0 = k_2 AA^*_{m-1} - k_2 AA_m^* - k_t AA_m^* - k_t' A_m^* S - k_D A_m^* C^* - k_C A_m^* C^*, \tag{32}$$

$$\frac{dS^*}{dt} = 0 = k_t' SC^* - k_s S^* A. \tag{33}$$

From these equations one can solve for A^*, A_m^* and S^*, respectively.

$$A^* = \frac{k_1 AB + k_t' SC^* + k_t AC^*}{k_2 A + k_t A + k_t' S + k_D C^* + k_C C^*}, \tag{34}$$

$$\frac{A_m^*}{A^*_{m-1}} = \frac{k_2 A}{k_2 A + k_t A + k_t' S + k_D C^* + k_C C^*} = r. \tag{35}$$

But $\quad C^* = A^* + A_2^* + A_3^* + \cdots = A^* + rA^* + r^2A^* + \cdots$

$$= A^* \sum_{n=0}^{\infty} r^n = \frac{A^*}{1 - r} \tag{36}$$

Substituting into equation (36) from equations (34) and (35)

$$C^* = \left(\frac{k_1 A B}{k_D + k_C}\right)^{1/2}. \tag{37}$$

This is the same result which in equation (18) was obtained from simpler considerations. Further work will be simplified by making the substitutions

$$\zeta = \frac{(k_1 B)^{1/2}\,(k_D + k_C)^{1/2}}{(k_2 + k_t)A^{1/2}}; \tag{38}$$

$$\beta = \frac{k_t}{k_2 + k_t}; \qquad \epsilon = \frac{k_D}{k_D + k_C}, \qquad \lambda = \frac{k_t' S}{(k_2 + k_t)A} \tag{39}$$

In terms of these parameters,

$$r = \frac{1 - \beta}{1 + \zeta + \lambda}; \qquad \frac{A^*}{C^*} = 1 - r = \frac{\zeta + \lambda + \beta}{\zeta + \lambda + 1} \tag{40}$$

Equation (29) reduces to

$$-\frac{dA}{dt} = (k_2 + k_t)AC^*\left[1 + \frac{(k_D + k_C)C^*}{(k_2 + k_t)A} - \frac{k_t A^*}{(k_2 + k_t)C^*} + \frac{k_t' S}{(k_2 + k_t)A} - \frac{k_D A^*}{(k_2 + k_t)A}\right], \tag{41}$$

$$= (k_2 + k_t)AC^*\left[1 + \zeta + \lambda - \beta\left(\frac{\zeta + \lambda + \beta}{\zeta + \lambda + 1}\right) - \epsilon\zeta\left(\frac{\zeta + \lambda + \beta}{\zeta + \lambda + 1}\right)\right] \tag{42}$$

But ζ is just the ratio of the rate of termination to the rate of propagation (the reciprocal of the kinetic chain length) which, as we have seen, is small. Likewise, β and λ were found to be the order of 10^{-3}. Therefore, the terms in the bracket will be negligible compared to unity if long chains are produced. This permits a simple change in variable from time, t, to monomer concentration, A, in equation (28), giving

$$\frac{dA_m}{dA} = \frac{dA_m}{dt} \bigg/ \left(-\frac{dA}{dt}\right)$$

$$\frac{k_D A_m}{(k_2 + k_t)A} + \frac{(k_t A + k_t' S)A_m^*}{(k_2 + k_t)AC^*} + \frac{1}{2}\,\frac{k_C \sum_{j=1}^{m-1} A_j^* A^*_{m-j}}{(k_2 + k_t)AC^*} \tag{43}$$

But $A_m^* = A^* r^{m-1}$, so that

$$-\frac{dA_m}{dA} = \frac{k_D}{(k_2 + k_t)}\,\frac{A^*}{A}\, r^{m-1} + \frac{(k_t A + k_t' S)A^*}{(k_2 + k_t)AC^*}\, r^{m-1} + \frac{1}{2}\,\frac{k_C A^{*2}(m-1)r^{m-2}}{(k_2 + k_t)AC^*} \tag{44}$$

From equation (40),

$$A^* = \frac{\zeta + \lambda + \beta}{\zeta + \lambda + 1}\, C^*,$$

so that equation (44) reduces to

$$\frac{dA_m}{dA} = (\epsilon\zeta + \lambda + \beta)\left(\frac{\zeta + \lambda + \beta}{\zeta + \lambda + 1}\right) r^{m-1} + \frac{1}{2}(m-1)(1-\epsilon)\zeta\left(\frac{\zeta + \lambda + \beta}{\zeta + \lambda + 1}\right)^2 r^{m-2} \tag{45}$$

This equation contains as special cases those distribution functions previously derived by probability considerations. Thus for $\beta = \lambda = 0$, $\epsilon = 1$, i.e., neglecting transfer and supposing termination to be by disproportionation, we have, after substituting for r from equation (40),

$$-\frac{dA_m}{dA} = \zeta^2\left(\frac{1}{1+\zeta}\right)^m \approx (1-\alpha)^2\alpha^m, \tag{46}$$

where $\alpha = 1 - \zeta$ is the probability of continuing the chain at any step. This is identical with the function derived by Dostal and Mark,[72] Flory[73] and Schulz[74] by a different method.[75] In the case $\beta = \lambda = 0$, $\epsilon = 0$, i.e., supposing termination to be by combination of free radicals, equation (45) reduces to

$$-\frac{dA_m}{dA} = \frac{1}{2}\zeta^3 m\left(\frac{1}{1+\zeta}\right)^m \approx \frac{1}{2}(1-\alpha)^3 m\alpha^m \tag{47}$$

---

[72] Dostal, H., & Mark, H. Zeit. physikal. Chem. B29: 299. 1935.
[73] Flory, P. J. Jour. Am. Chem. Soc. 58: 1877. 1936.
[74] Schulz, G. V. Zeit. physikal. Chem. B32: 27. 1936.
[75] See also, Gee, G. Trans. Faraday Soc. 31: 969. 1935, 32: 656, 666. 1936, and Marei, F. Acta Physicochim. URSS 9: 741, 759. 1938, 11: 549. 1939, who considered the case of termination by monomer.

which is identical with the modified function derived by Schulz[76] to account for the molecular weight distribution in polystyrene.

From equation (38), it is clear that in the case of first-order initiation, ζ is a function of monomer concentration. Consequently, equations (46) and (47) can represent the experimental molecular weight distribution only at the very beginning of the polymerization. For samples more than 5 per cent polymerized, these expressions must be integrated with respect to A before comparison with experiment is made. If the initiation is second-order in monomer and chain transfer can be neglected, then ζ is independent of A and equations (46) and (47) hold as they stand. This is conceivably the case in the uncatalyzed polymerization of styrene from solutions. However, in the absence of chain transfer, the mean molecular weight is also independent of the concentration of monomer. The experiments of Schulz[62] indicate, however, that the molecular weight increases as the reaction proceeds, and we have seen how the mechanism we have proposed for initiation, together with chain transfer reactions with the solvent are able to explain his data. The parameter, λ, which contains the rate of transfer to solvent, also contains A, the monomer concentration. Hence, if transfer is important, the distribution law will change as the reaction proceeds. Thus it is important that the per cent of monomer polymerized be known for every sample for which the molecular weight distribution function is to be determined.[77]

Schulz[78] has fractionated samples of polystyrene polymerized in solution without catalyst, as well as one sample of polystyrene polymerized as a pure liquid. Equation (47) was found by him to fit the observed distribution curves within the rather large experimental error for all the samples polymerized from solution. The pure styrene, however, appears to have a much sharper peak than equation (47) permits, as well as having a much higher maximum than can be obtained by any adjustment of the parameter. This sample was taken from a polymerizate which was 60 per cent polymerized, so that one should not expect the unintegrated form of the distribution function to apply.

For a first-order initiation reaction, we have from equation (38)

$$\zeta = \zeta_0 A_0^{1/2}/A^{1/2}, \tag{48}$$

where ζ_0 is the value of ζ for the initial concentration of monomer, A_0.

---

[76] Schulz, G. V. Zeit. physikal. Chem. B43: 25. 1939.

[77] As this was being written, a paper by Herington, E. F., & Robertson, A., appeared in the Trans Faraday Soc. 38: 490. 1942, which treats the molecular weight distribution function from the standpoint presented here, but without reference to any specific mechanism of chain initiation

[78] Schulz, G. V. Zeit physikal. Chem B43: 47. 1939.

Changing variable in equation (45) from A to ζ, and putting $\lambda = 0$, since there is no solvent, we obtain

$$\frac{dA_m}{d\zeta} = A_0\zeta_0^2 \frac{(\epsilon\zeta + \beta)(\zeta + \beta)}{\zeta^3} \frac{(1-\beta)^{m-1}}{(1+\zeta)^m} + A_0\zeta_0^2 \frac{(m-1)}{2} \frac{(1-\epsilon)}{\zeta^2} (\zeta + \beta)^2 \frac{(1-\beta)^{m-2}}{(1+\zeta)^m} \qquad (49)$$

Since we are interested primarily in values of $m > 100$, we can make the approximations

$$(1-\beta)^{m-1} \approx e^{-\beta m}$$
$$(1+\zeta)^{-m} \approx e^{-m\zeta}. \qquad (50)$$

The integration is then straightforward and gives

$$\frac{A_m}{A_0} = 2\zeta_0^2 e^{-\beta m} \left\{ \epsilon\left(1 - 2\beta m + \frac{\beta^2 m^2}{2}\right) \int_{\zeta_0}^{\zeta} \frac{e^{-m\zeta}}{\zeta}\, d\zeta + \left[\beta(1-\epsilon) - \frac{\beta^2 m\epsilon}{2}\right]\left[\frac{e^{-m\zeta_0}}{\zeta_0} - \frac{e^{-m\zeta}}{\zeta}\right] + \frac{\beta^2}{2}\left[\frac{e^{-m\zeta_0}}{\zeta_0^2} - \frac{e^{-m\zeta}}{\zeta^2}\right] + \left(\frac{1-\epsilon}{2}\right)\left[e^{-m\zeta_0} - e^{-m\zeta}\right] \right\}. \qquad (51)$$

The integral, $\int_{\zeta_0}^{\zeta} \frac{e^{-m\zeta}}{\zeta}\, d\zeta$, can be expressed in terms of the tabulated exponential integral[79]

$$\int_{\zeta_0}^{\zeta} \frac{e^{-m\zeta}}{\zeta}\, d\zeta = Ei(-m\zeta) - Ei(-m\zeta_0). \qquad (52)$$

Equation (51) gives the number distribution. The weight distribution is obtained by multiplying by the degree of polymerization, m.

This function has been applied to Schulz's data for the case of termination by combination of radicals, which seems to fit the experimental data slightly better. We have then $\epsilon = 0$ and

$$\frac{mA_m}{A_0} = m\zeta_0^2 \left[e^{-m\zeta_0}\left(1 + \frac{\beta}{\zeta_0}\right)^2 - e^{-m\zeta}\left(1 + \frac{\beta}{\zeta}\right)^2 \right] \qquad (53)$$

In FIGURE 7 are plots of $m^2 A_m/A_0$ versus m for different values of β/ζ_0 for a sample 60 per cent polymerized. From equation (39) we see that $\beta/\zeta_0 = \dfrac{k_t A C^*}{(k_D + k_C)C^{*2}}$, i.e., it is the probability that a growing polymer will

[79] Tables of sine, cosine and exponential integrals. WPA Computing Project. New York, N Y 1940.

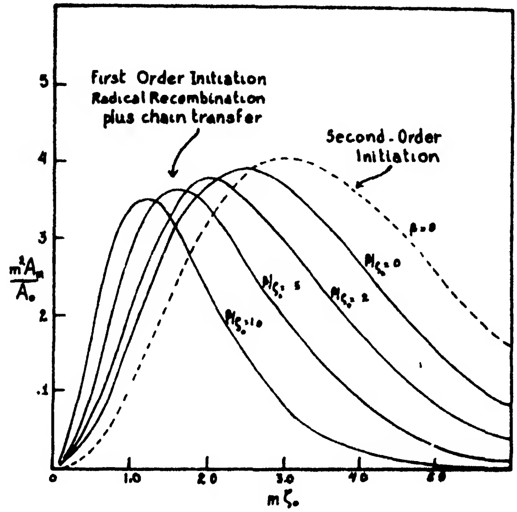


FIGURE 7. Theoretical molecular weight distribution for first order initiation calculated for 60 per cent polymerization

be terminated by transfer rather than by combination with another radical. Increasing the probability of transfer is seen to sharpen the distribution as well as to reduce the height at the maximum. FIGURE 8 shows the distribution reported by Schulz as well as that given by equation (53) for suitable values of ζ_0 and $\beta'\zeta_0$. No possible choice of values for these constants can reproduce the height of the maximum indicated by Schulz. This circumstance led to a re-examination of the method used to determine the molecular weights.

The "experimental" weight distribution curve is obtained by a graphical differentiation of the curve obtained by plotting weight per cent of polymer with molecular weight less than m versus m. The molecular weights were obtained by measuring the viscosities of dilute solutions of the polymer fractions and applying the Staudinger formula:

$$\frac{\eta_{sp}}{C} = K_m M, \tag{54}$$

where M is the molecular weight. K_m was determined by measuring the molecular weights of two of the fractions osmotically. There is evidence, however, that equation (54) is not obeyed by all substances.

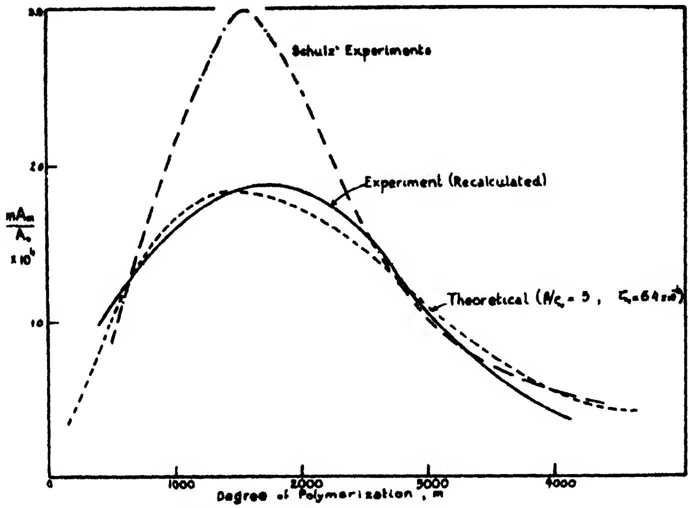


FIGURE 8 Molecular weight distribution in pure liquid styrene

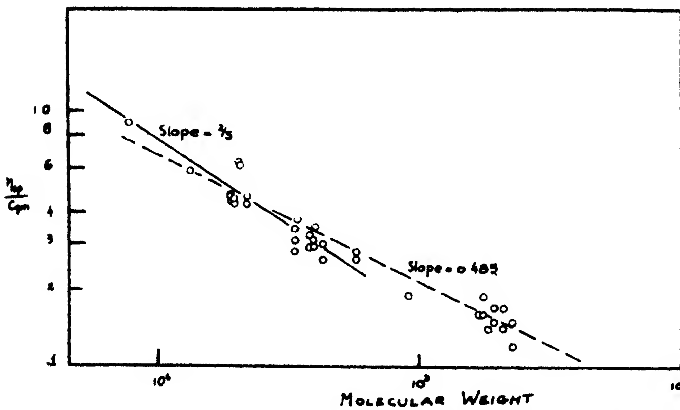


FIGURE 9 Viscosity of polystyrene as a function of molecular weight. Data from Staudinger (Ref 81)

Thus, in polyisobutylene,[80] the viscosity is proportional to the .640 power of the osmotic molecular weight. FIGURE 9 is a plot of the logarithm of molecular weight versus log (η_{sp}/c) for polystyrene, data

[80] Flory, P. J. Buffalo Meeting Am. Chem. Soc. September, 1942. Also Houwink, R. Jour prakt. Chem. 157: 15. 1940. for viscosity of polyesters.

from Staudinger.[31] Although the points scatter considerably, it is clear that the viscosity is far from proportional to the first power of molecular weight, in which case the line would be parallel to the 45°-line. For the region of high molecular weights, these data are much better represented by

$$\frac{\eta_{sp}}{C} = K_m' M^{2/3}. \tag{55}$$

It is, of course, not possible to fix the exponent closely from these data, and no particular significance should be given the number $^2\!/_3$ beyond the fact that it is an empirical result easy to compute with.

The molecular weights of the polystyrene fractions were recomputed from Schulz's data by equation (55) and the resulting weight distribution curve is labelled "recalculated" in FIGURE 8. This curve is much broader and lower than that given by Schulz and is easily fitted by the values $\zeta_0 = 6.4 \times 10^{-4}$, $\beta/\zeta_0 = 0.5$. The values $\zeta_0 = 10^{-3}$; $\beta/\zeta_0 = 0$ fit nearly as well. Thus, the data are not sufficiently accurate to permit a close calculation of the kinetic chain length, $1/\zeta$, or of the probability of chain transfer, β/ζ_0. However, the values indicated are very near those given in TABLE 2 obtained from the kinetic data, viz. $5.4 \times 10^{-4}/A_0^{\frac{1}{2}} \approx 1.8 \times 10^{-4}$. A more reliable determination of the molecular weights would make possible a much more precise determination of these constants.

For the case of second-order initiation, ζ is independent of monomer concentration, and equation (45) must be integrated with respect to λ to account for the effect of chain transfer in solvents. Thus we have, since $\lambda = \lambda_0 A_0/A$,

$$\frac{dA_m}{d\lambda} = \frac{\lambda_0 A_0}{\lambda^2} e^{-m(\beta+\lambda+\zeta)} \left[(\epsilon\zeta + \lambda + \beta)(\zeta + \lambda + \beta) + \frac{m}{2}(1 - \epsilon)\zeta(\zeta + \lambda + \beta)^2 \right], \tag{56}$$

[31] Staudinger, H. "Die Hochmolekulare Organische Verbindungen." Julius Springer. Berlin. 1932.

$$\frac{A_m}{\cdot} = \lambda_0 e^{-m(\beta+\zeta)} \quad (\zeta + \beta)(\epsilon\zeta + \beta) +$$

$$\left(\frac{1-\epsilon}{2}\right) m\zeta(\zeta + \beta)^2 \Bigg] \left[\frac{e^{-m\lambda_0}}{\lambda_0} - \frac{e^{-m\lambda}}{\lambda}\right]$$

$$+ \Bigg[(1 + \epsilon)\zeta + 2\beta + m\zeta(1 - \epsilon)(\zeta + \beta) - \beta^2$$

$$- \left(\frac{1-\epsilon}{2}\right) m\zeta(\zeta + \beta)^2 \Bigg] \Bigg[Ei(-m\lambda_0) - Ei(-m\lambda) \Bigg]$$

$$+ \left[1 + \left(\frac{1-\epsilon}{2}\right) m\zeta \right] \left[e^{-m\lambda_0} - e^{-m\lambda} \right] \Bigg\} \qquad (57)$$

For $\beta = \epsilon = 0$, i.e., when there is no transfer directly to monomer and termination is by combination of growing polymer radicals,

$$\frac{A_m}{A_0} = \lambda_0 e^{-m\zeta} \left\{ \frac{1}{2} m\zeta^3 \left(\frac{e^{-m\lambda_0}}{\lambda_0} - \frac{e^{-m\lambda}}{\lambda}\right) \right.$$

$$+ \left(\zeta + m\zeta^2 - \frac{1}{2} m\zeta^3\right) [Ei(-m\lambda_0) - Ei(-m\lambda)]$$

$$\left. + \left(1 + \frac{m\zeta}{2}\right) (e^{-m\lambda_0} - e^{-m\lambda}) \right\} \qquad (58)$$

The equation is similar in form to equation (51) except that the roles of λ and ζ are interchanged. Thus, the form of the distribution curve is not sufficiently sensitive to the kinetics of the polymerization to establish the difference between first- and second-order initiation or the extent of transfer to the solvent from the existing data. Distributions based on a larger number of fractions might make a decision possible on some of these points.

Number Average Molecular Weight

The effect of chain transfer on the molecular weight is given by equations (21) and (23) during the initial stages of the reaction. However, as the monomer becomes depleted, it is evident that the molecular weight of the product molecules decreases, so that shorter chains are formed at the end of the reaction than at the beginning. The mean molecular weight of the polymerizate at high percentage conversion is then not given by these equations. The number average molecular weight is defined to be

$$\bar{P}_N = \frac{\sum_{m=2}^{\infty} mA_m}{\sum_{m=2}^{\infty} A_m}, \tag{59}$$

where A_m is the concentration of m-mer. The sum in the numerator is just the total weight of polymer, i.e., $(A_0) - (A)$, at any given monomer concentration (A). The sum in the denominator can be evaluated most easily from equation (45). We have

$$\frac{d}{dA}\sum_{m=2}^{\infty} A_m = \sum_{m=2}^{\infty}\frac{dA_m}{dA},$$

$$-\sum_{m=2}^{\infty}\frac{dA_m}{dA} = (\epsilon\zeta + \lambda + \beta)\left(\frac{\zeta + \lambda + \beta}{\zeta + \lambda + 1}\right)\sum_{m=2}^{\infty} r^{m-1}$$
$$+ \left(\frac{1-\epsilon}{2}\right)\zeta\left(\frac{\zeta + \lambda + \beta}{\zeta + \lambda + 1}\right)^2 \sum_{m=2}^{\infty}(m-1)r^{m-2}. \tag{60}$$

The summations are easily made and we have

$$-\sum_{m=2}^{\infty}\frac{dA_m}{dA} = \left(\frac{1+\epsilon}{2}\right)\zeta + \lambda + \beta. \tag{61}$$

For first-order initiation, $\zeta = \zeta_0 A_0^{1/2}/A^{1/2}$ and $\lambda = \lambda_0 A_0/A$. Making these substitutions and integrating gives

$$\sum_{m=2}^{\infty} A_m = (1+\epsilon)\zeta_0 A_0^{1/2}(A_0^{1/2} - A^{1/2}) + \lambda_0 A_0 \ln\frac{A_0}{A} + \beta(A_0 - A), \tag{62}$$

$$\frac{1}{\bar{P}_N} = \beta + \frac{\lambda_0 A_0}{A_0 - A}\ln\frac{A_0}{A} + \frac{(1+\epsilon)\zeta_0 A_0^{1/2}}{A_0^{1/2} + A^{1/2}} \tag{63}$$

For small values of $\frac{A_0 - A}{A_0}$ (small conversions), this reduces to equation (21).

For second-order initiation, ζ and β are constants, and the integration gives

$$\sum_{m=2}^{\infty} A_m = \left[\left(\frac{1+\epsilon}{2}\right)\zeta + \beta\right][A_0 - A] + \lambda_0 A_0 \ln\frac{A_0}{A}, \tag{64}$$

$$\frac{1}{\bar{P}_N} = \left(\frac{1+\epsilon}{2}\right)\zeta + \beta + \frac{\lambda_0 A_0}{A_0 - A}\ln\frac{A_0}{A}, \tag{65}$$

which reduces to equation (23) for small values of $\frac{A_0 - A}{A_0}$.

Radical Initiation

A third type of initiation occurs where free radicals are introduced into the system. Thus, addition to styrene of tetraphenyl succinic acid dinitrile, which decomposes to give Gomberg-type free radicals, accelerates the polymerization.[82] Triphenyl methyl azobenzene acts in a similar fashion. We have already considered the possibility that benzoyl peroxide catalyzes styrene polymerization by an initial decomposition into radicals. Introduction of free radicals from decomposing metal alkyls has been shown to initiate polymerization,[83] as has the action of decomposing azomethane. There can be no doubt that free radicals, if present, can and do induce polymerization. This occurs, presumably, by the addition of the radical to the monomer to form a polymer radical capable of growth. A possible kinetic scheme has already been presented in the discussion of styrene polymerization.

Conclusions

From the standpoint of elucidating the mechanism of polymerization, it would be very desirable to measure rates in dilute solutions of monomer, since these data are more readily analyzed theoretically. Moreover, it is very difficult to determine reaction order accurately from data at a single initial concentration of monomer. Especially are these data needed for the acid catalyzed polymerizations. Molecular weight distribution curves, when sufficiently accurate, can be used to confirm the mechanism adopted on the basis of kinetic experiments, and, in addition, can furnish a decision as to the nature of the termination process. Improvement in the methods of determining mean molecular weights is necessary. The viscosity law, apparently, needs to be established experimentally for each polymer in each different solvent if it is to give reliable molecular weights.

In spite of these imperfections in the experimental data, a fairly consistent interpretation of the kinetics of styrene polymerization is possible based on the hypothesis that a low-lying triplet state in ethylene is capable of entering into the activated complex in the chain initiation process. Whether or not this is the correct interpretation remains to be seen, but, at any rate, it is one worthy of consideration.

---

[82] Schulz, G. V. Zeit. Elektrochem. 47: 265. 1941.
[83] Taylor, H. S., & Jones, W. H. Jour. Am. Chem. Soc. 52: 1111. 1930.

B. PHYSICAL PROPERTIES OF HIGH POLYMERS

Polymeric materials in the rubbery state of aggregation are characterized both by a long range and a local structure. The long range structure is a polymeric network, the network units being composed of sparsely-occurring bonds between the coiled polymeric chains. These bonds may be primary cross links, or secondary bonds such as dipole–dipole bonds, or regions of local crystallinity. The local structure is very similar to the local structure of liquids, the segments of the long chain molecules being in continual, fairly rapid motion, moving from one equilibrium position to the next.[84] It is desirable to construct a theory of the over-all physical properties based on molecular quantities such as the number of bonds of a given type per unit volume, the strength of these bonds, and the internal viscosity of the local liquid-like structure.[85]

Characterization of a Network

The formation of three dimensional structures by cross linking has been discussed by several authors.[86] Incipient "gel" formation according to Flory starts when there is one cross link per two chains. Materials that are in a state comparable to soft vulcanized rubber have this three dimensional structure, as witnessed by the fact that they swell rather than dissolve in likely solvents, and because of their highly retractive elastic behavior.

The structure is built up from cross linking between linear chains. One very important quantity characterizing the network is ν, the number of network junctures per unit volume. Another important quantity is l, the mean distance between neighboring network bonds projected in a given direction.

It is often clear that the network juncture bonds divide into definite classes, such as when primary cross links and secondary cross bonds are present. In this case it is convenient to indicate the i-th type of bond by the subscript i and define quantities ν_i and l_i.

Let us suppose that the stress f_i has distributed itself on the i-th kind of bonds. We wish to determine what is the average force that acts on each bond of type N. It can be seen that this is N_i, where $N_i = \nu_i l_i$.

---

[84] References on the structure of polymeric materials include Mark, H. Ind. Eng. Chem. 34: 449. 1942; Mark, H. Cold Spring Harbor Symposia on Quantitative Biology, 9: 204. 1942; Alfrey, T., & Mark, H. Jour. Phys. Chem. 46: 112 1942; Leaderman, H. Lecture presented at Am. Chem Soc. meeting in Buffalo, N. Y., September, 1942; Scott, J. R. Trans. Faraday Soc. 38: 284. 1942. Kauzmann, W., & Eyring, H. Jour. Am. Chem. Soc. 62: 3113. 1940.

[85] Tobolsky, A., & Eyring, H. Jour Chem. Phys. 11: 125 1943.

[86] Flory, P. J. Jour. Phys. Chem. 46: 132–140. 1942.

We shall call n the number of bonds per unit area. Then if we call $n_i = \frac{1}{l_i}$ the number of bonds per unit length, we have $\nu_i = n_i N_i$.

The Statistics of Long-chain Molecules

Having discussed the characterization of these polymeric networks, it is necessary to point out that the mobile chains between the network junctures can and do assume many conformations in consequence of the thermal agitation. In discussing the elastic properties of these substances it turns out that it is of importance to know just how many ways a coiling molecular chain can assume a length x between its ends (or between juncture points in a lattice). Let us for simplicity assume that we are dealing with a hydrocarbon chain. The stable configuration for successive bond directions is believed to be the staggered configuration. Then the meanderings of the hydrocarbon chain can be described as a continuous path on a diamond lattice, each path being equally likely if we neglect the effect of steric hindrance.

Suppose that the total number of bonds is n and the bond length divided by $\sqrt{3}$ is taken as the unit of length. The random walk on a diamond lattice problem turns out to be equivalent to the problem of finding the distribution in heads of a penny that remembers its last flip and has a $2/3$ chance of repeating that flip.

The result of a rather lengthy calculation is that the probability of a chain of n bonds having the length x is given by[87]

$$P_{n\,(x)} = \frac{1}{2\pi^{1/2}n^{1/2}} e^{(-x^2/4n)}. \tag{66}$$

This formula provides a rigorous basis for the discussion of entropy elasticity of rubber-like materials.

Rubber-like Elasticity

From thermodynamics we can immediately derive the law

$$\tau = \left(\frac{\partial E}{\partial l}\right)_T - T\left(\frac{\partial S}{\partial l}\right)_T, \tag{67}$$

where τ is tension and l represents length. Various considerations point to the fact that the term involving change of entropy with length is of great importance. For example the tension at constant extension has been claimed to be proportional to the absolute temperature. Also,

[87] Tobolsky, A., Powell, R. E., & Eyring, H. Chapter on "Elasto-Viscous Properties of Matter" in "Frontiers in Chemistry." Interscience Publishers. P 125. 1943.

there is a temperature rise on rapid stretching of the rubber. However, the evidence is that both internal energy, E, and entropy, S, change with an uncoiling of the molecular chains.

The entropy elasticity can be calculated in terms of the statistics of long chain molecules.[87, 88] Let us suppose that there are altogether Z independent network units in a macroscopic piece of rubber. From the statistics of long-chain molecules it can be seen that the probability of a network unit having the dimensions x, y, z, is

$$\mathrm{P}(xyz) = Ae^{-\beta(x^2+y^2+z^2)}. \tag{68}$$

The network units are, to a large extent, interpenetrating, so that we can assume that they have a constant volume, v.

The equation of state can be derived from the principle of Boltzmann relating entropy to probability, $S = k \ln \mathrm{P}$ and from the relation between tension and entropy, namely $\tau = -T\left(\dfrac{\partial S}{\partial l}\right)_T$. If one further assumes that $l = l_0 \dfrac{x}{v^{1/3}}$, where l_0 is the unstretched length, then the following equation of state can be derived:

$$\tau = 2\,\frac{\beta v^{2/3} ZkT}{l_0}\left(\frac{l}{l_0} - \left(\frac{l_0}{l}\right)^2\right). \tag{69}$$

The modulus of elasticity is therefore

$$G = C'\nu kT, \tag{70}$$

where C'' is a constant of the order of magnitude of unity and ν is the number of bonds per unit volume.

It is to be realized that the characteristic network units are closed loops formed by cross links between chains.

Unit Process of Deformation

Consider a kinetic unit of a condensed phase at rest in a position of equilibrium at point A, but having the possibility of surmounting a free energy barrier and attaining a new position at point B. When no stress is acting, the number of times the kinetic unit jumps forward is equal to the number of times that it jumps back, so there is no net motion. However, when stress is applied to the medium there is a twofold action. In the first place there is an elastic displacement in phase with the stress, and secondly, the stress favors motion of the kinetic unit from one posi-

[88] Guth, E., & James, H. Ind. Eng. Chem. 33: 624. 1941. See earlier work of Meyer, Mark, Kuhn, and Guth referred to in this paper and in reference 87.

tion of equilibrium to the next. Mathematically formulated, the stress is added as a linear potential to the expression for the free energy as a function of distance. The elastic displacement then occurs because the molecule has shifted its position of equilibrium in the free energy well, and the flow term occurs because the free energy of activation is lowered in one direction and raised in the other.

For a cosine-shaped barrier it can be shown that the rate of strain is expressed in terms of the stress[87] as follows:

$$\frac{ds}{dt} = \frac{1}{G}\frac{df}{dt} + \frac{\lambda}{\lambda_1}\frac{kT}{h}e^{-(\Delta F^{\ddagger}/kT)}\,2\sinh f\frac{\lambda\lambda_2\lambda_3}{2kT}, \tag{71}$$

where G is the elastic modulus, λ the distance between equilibrium positions, $\lambda_2\lambda_3$ the area of the kinetic unit in the plane of stress. $\Delta F^{\ddagger}$ is the height of the free energy barrier and G is related to the curvature in the free energy well. The moduli associated with the local structure in all condensed phases is of the order of 10^{11} dynes/cm^2 because of the nature of the forces between atoms.

Equations of Motion for Network Structures

The equations derived above are suitable for substances composed of small molecules, and the flow term has been applied in interpreting liquid viscosities. For network structures, it is necessary to modify the equations so as to introduce the quantities ν, N, and n defined in a previous section. Cross bonds of type i give rise to two terms in the rate of strain, an elastic term referring to the uncoiling of long chain molecules between these cross bonds, and a term involving the contribution of slipping and breaking at the cross bonds. The equation for the i-th type of network unit is[85]

$$\frac{ds}{dt} = \frac{1}{G_i}\frac{df_i}{dt} + n_i\lambda_i\frac{kT}{h}e^{-(\Delta F^{\ddagger}/kT)}\,2\sinh\frac{f_i\lambda_i}{2N_ikT}, \tag{72}$$

where, as before, if l_i is the mean distance between bonds of type i projected in the direction of stress, and ν_i is the number of such bonds per unit volume, then $n_i = 1/l_i$ and $N_i = l_i\nu_i$. f_i is the stress distributed on the i-th kind of network bonds and λ_i is the distance between equilibrium positions.

An important question arises here, namely: Can the distance λ_i be correlated to l_i, so that $n_i\lambda_i$ is of the order of magnitude of unity, or is it true that the distance traversed in the unit process is of the order of magnitude of a few angstroms, as in the case of liquids? If λ_i is of the

same order as the distance between network bonds, then the equation of motion can be written completely in terms of ν_i, the number of network bonds of type i.

$$\frac{ds}{dt} = \frac{1}{C'\nu_i kT}\frac{df_i}{dt} + \frac{kT}{h} e^{-(\Delta F^{\ddagger}/kT)}\, 2 \sinh \frac{f_i}{2\nu_i kT} \tag{73}$$

where $c'\nu_i kT$ is substituted for G_i. Under suitable assumptions, it can be shown that the formula presented in an earlier section for entropy elasticity gives rise to a modulus of $6\nu_i kT$, as was shown in detail by Flory at this meeting. It must be remembered, however, that such is the case only if there is no change in internal energy on stretching, so that in many cases the modulus may be much bigger than the formula indicates.

Combination of Unit Flow Processes

In order to arrive at equations of motion for the medium it is necessary to make some postulate as to the way the stress distributes itself on the unit processes of deformation, and how the motion of these units is reflected in the total motion of the medium. The most reasonable assumption is that the stress distributes itself on the unit processes in such a way that the rate of strain is the same for each of them

Most of the interesting mechanical behavior of polymeric materials can be explained by assuming that three structural elements of the network must be considered, namely: network units which are held together by primary cross-linking bonds, network units held together by secondary cross-linking bonds, and the motion of the mobile chain segments. The following equations of motion are appropriate·

primary network units

$$\frac{ds}{dt} = \frac{1}{G_1}\frac{df_1}{dt} + n_1\lambda_1 \frac{kT}{h} e^{-(\Delta F^{\ddagger}/kT)}\, 2 \sinh \frac{f_1\lambda_1}{2N_1 kT} \tag{74}$$

secondary network units

$$\frac{ds}{dt} = \frac{1}{G_2}\frac{df_2}{dt} + n_2\lambda_2 \frac{kT}{h} e^{-(\Delta F_2^{\ddagger}/kT)}\, 2 \sinh \frac{f_2\lambda_2}{2N_2 kT} \tag{75}$$

segment motion

$$\frac{ds}{dt} = \frac{1}{G_3}\frac{df_3}{dt} + \frac{1}{\eta_3} f_3. \tag{76}$$

Since very small stress is operative on the individual units of the local structure (i.e. the segments of the long-chain molecules), the

hyperbolic sine can be expanded, giving an equation such as written for the segment motion, which is of a form proposed empirically by Maxwell.[89–91] G_1 and G_2 are entropy moduli of the order of 10^7 dynes/cm^2 whereas G_3 is of the order of 10^{11} dynes/cm^2. The total stress is, of course,

$$f = f_1 + f_2 + f_3.$$

We shall now discuss experiments that can be interpreted in terms of these equations of motion.

Stress Relaxation at Constant Elongation

If a polymeric thread be rapidly stretched to a given length, and held fixed at that length, the stress necessary to maintain that length decays as a function of time. For soft vulcanized rubber, Meyer[92] believed that there was an initial decay of stress and that finally a limiting stress was reached. However, Phillips[93] in very early experiments observed that the tension in a stretched band of India rubber appeared to continually relax, and that after a year at room temperature the tension was approaching zero. He also observed that initially the stress was a linear function of log time, but that after a time the decay to zero (on a logarithmic time plot) became more rapid.

From the equations of motion and the model of the molecular structure, this behavior can be understood. In the initial rapid stretching, neither primary nor secondary bonds have time to slip, so that the stress distributes on these bonds. Thereafter, bonds begin to slip and reform, the weaker bonds first and the stronger after long periods of time. The slipping of bonds allows the long chains to coil into random rather than stretched configurations, and the stress is thereby released.

Neglecting f_3, the total stress is $f_1 + f_2$. Both of these decay according to the law

$$\tanh \frac{f_i \lambda_i}{4N_i kT} = \tanh \frac{f_i^0 \lambda_i}{4N_i kT} e^{-k_i' t}, \tag{77}$$

$$\text{where } k_i' = \frac{G_i n_i \lambda_i^2}{2N_i kT} \frac{kT}{h} e^{-(\Delta F_i^{\ddagger}/kT)}. \tag{78}$$

If $n_i\lambda_i$ is of the order of unity, k_i' is approximately three times the

[89] For mathematical formulation of the properties of elastoviscous bodies using Maxwell's equation and distributions of relaxation times see references 89, 90, 91. Kuhn, W. Zeit. physik. Chem. 42: 1. 1939; Bennewitz, K., & Rötger, H. Physik. Zeits. 40: 416. 1939.
[90] Simha, R. Jour. Appl. Phys. 12: 630. 1941
[91] Ferry, J. D. Jour. Am. Chem. Soc. 64: 1330. 1942.
[92] Meyer, K., & Ferri, C. Helv. Chim. Acta 18: 574. 1935.
[93] Phillips, P. Proc. Phys. Soc. 19: 491. 1905.

natural relaxation frequency of the bonds, and the relaxation equation becomes

$$\tanh \frac{f_i}{4\nu_i kT} = \tanh \frac{f_i^0}{4\nu_i kT} e^{-k_i' t}. \tag{79}$$

If the stretching process involves no change in internal energy, $f_i^0 = 2\beta v^{\frac{1}{2}} \nu_i kT\gamma$ (equations [69], [70]) where γ is a function of the elongation. However, f_i^0 may be considerably larger than this if internal energy is changed.

If $f_i^0 > f > 4\nu_i kT$, f_i plotted against log time will be linear. If on the other hand, $f_i < 4\nu_i kT$ then Maxwell's equation for the relaxation of stress will apply, namely, $f_i =$ const. $e^{-k_i' t}$.

A good approximate value for the number of secondary bonds per unit volume can be obtained by setting the initial slope of the stress–log time equal to $2\nu_2 kT$. The onset of the relaxation of stronger bonds should be indicated by a steeper slope. The energy of the bonds that are breaking are obtainable from the time at which the relaxation of stress due to the breaking of these bonds becomes appreciable. (See reference[85] for graphs.)

Inasmuch as secondary bonds are formed by interaction between the long-chain molecules, it is not surprising to find that the number of such bonds depends on elongation and temperature.

Other Experiments Revealing the Nature of the Network Structure

Another method for studying the strength and number of network bonds is to observe the behavior of these polymeric materials under constant stress. For these cases there is observed an initial deflection, followed by a creep. This creep becomes slower and for sufficiently small loads and small times a final limiting deflection appears to be reached if deflection is plotted against linear time. On a log time plot, much the same kind of behavior is observed as in stress relaxation. There is an initial linear slope followed by a steeper slope until finally the sample breaks (if the experiment is carried out for a sufficiently long time).

The initial deflection is, of course, determined by the number of primary and secondary network bonds. The initial creep occurs because the secondary bonds start breaking and reforming in new positions, while the primary cross links have not yet started relaxing to any appreciable extent. The rapid creep before breaking is due to the breaking of primary cross links. The mathematical formulation of the creep curve in terms of the equations of motion presented here is given by Tobolsky and Eyring.[85]

Polymeric networks with a large amount of primary cross links cannot be extruded through ordinary machines. If the network bonds are mainly secondary bonds, these substances can be extruded, but the extrusion rate is not a linear function of the pressure as is the case for Newtonian liquids which obey Poiseuille's equation. The equations of motion used here give the following equation,[87]

$$\frac{dV}{dt} = \frac{2\pi l R^2 A}{Bp}\left[\cosh\frac{BpR}{2l} - \frac{4l}{BpR}\sinh\frac{BpR}{2l} + \frac{16l^2}{(BpR)^2}\sinh^2\frac{BpR}{4l}\right], \quad (80)$$

where $\frac{dV}{dt}$ is the volume extruded per second, p is the extruding pressure, R the radius and l the length of the orifice, $B = \frac{\lambda_2}{2N_2kT}$ and $A = n_2\lambda_2\frac{kT}{h}e^{-(\Delta F_2^\ddagger/kT)}$ In FIGURE 10 are shown data obtained by Dillon and Johnson[94] and calculated curves obtained by adjusting the

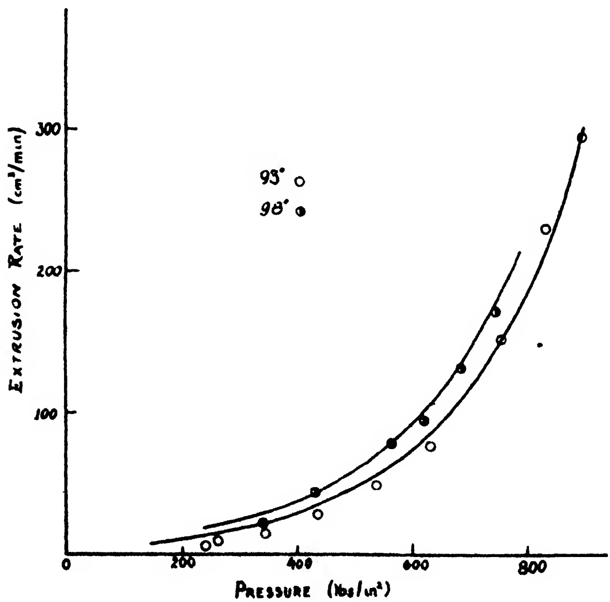


FIGURE 10 Extrusion of rubber.

[94] Dillon, J. H., & Johnson, N. Jour. Appl Phys 4: 225 1933.

two parameters A and B. From this data it appears that $\Delta F_2 \cong 15.5$ kcal.

Breaking under load can also be interpreted in terms of this model and the empirical equation relating the life-time τ of a thread holding a suspended weight W, namely,[95]

$$\log \tau = -aW + b,$$

where a and b are constants from which the strength and number of bonds can be determined.[95]

Effect of Local Structure on Physical Properties

The motion of the mobile chain segments has a damping effect on the mechanical properties of the network much as if a viscous liquid had been impregnated in the polymeric network. The nature of this internal viscosity can be revealed by high-frequency mechanical and electrical experiments. The relaxation times for the motion of these segments for rubbery substances are of the order of 10^{-6} seconds at room temperature compared to relaxation times of the order of seconds or hours for secondary bonds, and years for primary bonds. High-frequency experiments too rapid to catch the relaxation of secondary bonds will yield values of the internal viscosity of 10^3–10^5 poises.[96] This is to be compared to the very high viscosities that are measured by experiments such as extrusion, in which the resistance to flow is due to the necessity of breaking network bonds.

Long Molecules in Solution

When dissolved in a solvent with which there is no heat of solution, linear molecules will from probability considerations tend to assume the form assumed by a random chain. Recently an interesting statistical treatment of the extent of a random chain (that is, the smallest cubical box that will entirely contain the chain) has been carried out.[97] The result is that the most probable extent (projected in a given direction) of a chain of N links, each of which are of length a, is $1.32a\sqrt{N}$. This will then tell us the most probable volume occupied by long polymeric molecules in solution.

Kuhn[98] had treated the viscosity of long-chain molecules in dilute solution, assuming that each molecule together with the liquid that it entraps can be considered as forming an immobilized particle, for which

---

[95] Busse, W. F. Jour. Appl. Phys. 13: 1942, Busse, W. F., Lessig, E. T., Loughborough, D. L., & Larrick, L. Jour. Appl. Phys. 13: 715. 1942.
[96] Stambaugh, R. B. Ind. Eng. Chem. 34: 1362. 1938.
[97] Daniels, H. E. Proc. Cambridge Phil. Soc. 37: 244. 1941
[98] Kuhn, W. Kolloid Zeit. 68: 2. 1934.

the Einstein treatment of viscosity of suspensions is applicable. Using the statistical *extent* of chain, rather than the distance between ends of the chain as a measure of the amount of immobilized fluid, Kuhn's treatment would lead to the following equation for the specific viscosity

$$\eta_{sp} = \frac{12.7}{m^{3/2}} \gamma^{3/2} M^{1/2} C \times 10^{-3} \tag{81}$$

where m is the average molar weight of each carbon atom plus its attached side groups, M the molecular weight of the chain, γ the number of carbon atoms per freely orienting segment, and c the concentration in grams per liter. This, as will be seen, deviates considerably from the Staudinger law for which $\left(\frac{\eta_{sp}}{C}\right)_{\lim}$ is proportional to M. However, at least for the case of polystyrene, the 0.5 power law holds better than the first power using Staudinger's own data (see FIGURE 9) and the value $\gamma = 1.5$ is reasonable. For other substances, empirical laws involving M to a power between 0.5 and 1.0 have been proposed. It would appear that more exact and extensive data correlating intrinsic viscosity with molecular weight are needed, as well as a more complete theory.

The concentration at which the "extent cubes" around each molecule completely fill the volume of the solution is

$$C \times 10^{-3} = \frac{m^{3/2}}{5.18\gamma^{3/2} M^{1/2}}. \tag{82}$$

For polyethylene of molecular weight 100,000, and $\gamma = 1$, this critical concentration is 32 gm./liter. At any larger concentration the molecules would, therefore, start to interpenetrate, and new viscosity laws for this region are needed.

It is to be noted that in these equations, γ is the number of carbon atoms per segment, where the segments are freely orientable units from the point of view of calculating the statistical extent of chain, and do not necessarily have as large a length as the segment length calculated from viscosity measurement on molten polymers,[99] osmotic pressure curves and so on. For high molecular weights where branching becomes important, M in equation (81) becomes larger.

[99] Kauzmann, W., & Eyring, H. Jour Am. Chem Soc 62: 3113. 1940.

STATISTICAL THEORY OF CHAIN CONFIGURATION AND PHYSICAL PROPERTIES OF HIGH POLYMERS

By

Paul J. Flory and John Rehner, Jr.

From the Esso Laboratories, Chemical Division, Standard Oil Development Company, Linden, N. J.

INTRODUCTION

Polymeric molecules composed of long, primary valence chains can assume numerous configurations made possible by quasi-free rotation about single carbon to carbon bonds within the chain skeleton. The development of a statistical mechanical theory of chain configuration, based on the pioneering work of Eyring,[1] Guth and Mark[2] and Kuhn[3] has been particularly successful in explaining rubber-like elasticity of a number of substances composed of very long-chain molecules. In the undeformed state the chains assume configurations approximating closely to their most probable configuration. Deformation results in a displacement of chain configurations. Thus deformation is accompanied by a decrease in entropy, which is the major factor responsible for rubber-like elasticity.

Another phenomenon closely related to elastic deformation is that of limited swelling of cross-linked, or vulcanized, high polymers in contact with solvents. The network structure is expanded by the dissolved solvent up to the point where the elastic forces in the chains of the network become counterbalanced by the tendency toward dilution by solvent.

Various theories of chain configuration and their application to the deduction of relations between elastic properties and structure will be reviewed in the course of the following discussion. Special attention will be devoted to the effects of low chain flexibility due to hindered rotation about single bonds of the chain. A new model for network polymer structures which provides an alternate procedure for deriving elastic characteristics will be described.

CHAIN CONFIGURATION DISTRIBUTION

The probability distribution of the distances between the ends of a set of long chain molecules is of foremost importance in the treatment of problems relating to the deformability of high polymers, particularly

[1] Eyring, H. Phys. Rev. 39: 746. 1932.
[2] Guth, E., & Mark, H. Monatshefte für Chemie 65: 63. 1934.
[3] Kuhn, W. Kolloid-Z. 68: 2. 1934, 76: 258. 1936.

those possessing network structures. Guth and Mark[2] and Kuhn[3] concluded that the probability of a length between r and $r + dr$ for the vector connecting the ends of the chain, which is free to assume irregular configurations at random, is given by

$$W(r)dr = (4\beta^3/\pi^{1/2}) \exp(-\beta^2 r^2) r^2 dr, \tag{1}$$

where β is a parameter, the value of which will depend on the length of the chain and its flexibility. In Cartesian coordinates

$$W(x\ y\ z)dx\ dy\ dz = (\beta^3/\pi^{3/2}) \exp\left[-\beta^2(x^2 + y^2 + z^2)\right]dx\ dy\ dz, \tag{1'}$$

where the origin of coordinates is taken at one end of the chain and x, y, z, $dx\ dy\ dz$ defines the volume element containing the terminus of the vector leading to the other end of the chain.

The relative numbers of chains of various lengths r ("displacement lengths") are shown graphically in FIGURE 1. The most probable length r_{max}, the average and the root mean square lengths are given by

$$r_{max} = 1/\beta, \tag{2}$$

$$\bar{r} = \sqrt{4/\pi}/\beta, \tag{3}$$

$$\sqrt{\overline{r^2}} = \sqrt{3/2}/\beta. \tag{4}$$

Eyring[1] has derived a general relationship between $\overline{r^2}$, and the number Z of bonds in the chain, the length l of each bond, and angle ω between successive bonds. For tetrahedral bond angles ($\omega = 109.5°$) and free rotation about each bond, Eyring's relationship reduces to[2,3]

$$\overline{r^2} = 2l^2 Z. \tag{5}$$

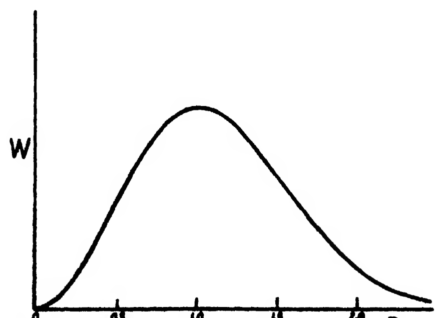


FIGURE 1. Probability distribution of distances r between the ends of long randomly kinked chains, r expressed in units of $1/\beta$, $W(r)$ in relative units. See equations (1) and (6).

Hence, for the case of free rotation

$$\beta = \sqrt{3/4Zl^2}. \tag{6}$$

Upon substituting equation (6) in (1′), Kuhn[3] and Guth and Mark[2] were able to express the probability distribution of chain displacement lengths r as a function of the chain contour length, Zl (or $Zl \sin \omega/2$).

If rotation about the chain bonds is accompanied by a *symmetrical* hindrance potential, then according to Eyring and co-workers[4] equation (5) requires no modification. Such a symmetrical hindrance potential is represented by

$$U = U_0(1 - \cos n\theta)/2, \tag{7}$$

where θ is the angle of bond rotation and n is an integer. The angle θ_i of rotation about the i-th bond is defined as the angle between bond $i + 1$ and its projection in the plane defined by bonds $i - 1$ and i. If, in analogy with the situation in the ethane molecule, there are three *equal* minima ($n = 3$) associated with rotation about each bond, the distribution of chain displacement lengths as defined by equations (1), (5), and (6) is unaffected by the hindrance potential.

Steric effects in long polymer molecules are certain to disrupt the symmetry of the bond rotation hindrance potential. This dissymmetry is likely to be accentuated when large substituent groups are attached to the chain. Recently Bunn[5] has called attention to the tendency in long-chain molecules (e.g., polyethylene, rubber, and polyesters) for the bonds to assume a "staggered configuration," the equivalent of the planar zig-zag structure having θ equal to zero. In other words, $\theta = 0$ frequently represents the lowest minimum in the bond rotational potential. Sequences of bonds in the planar zig-zag form should occur in such cases The average length of these sequences will depend on the depth of the potential minimum for $\theta = 0$, and on its depth as compared with minima for other values of θ. Such a chain would consist of numerous straight (planar zig-zag) sequences of bonds joined together in an essentially haphazard fashion.

The above analysis represents an oversimplification. In any actual case, the hindrance potential for any given bond will depend on the arrangement of its neighbors in a complicated manner. Without attempting any detailed analysis, Kuhn[3] considers that hindrance of rotation about chain bonds necessitates replacement of the single bond of length l by a larger unit composed of s bonds as the freely orienting segment. Z must then be decreased by the factor $1/s$ and l must be replaced by the

[4] Gorin, E., Walter, J., & Eyring, H. Jour. Amer. Chem. Soc. 61: 1885. 1939
[5] Bunn, C. W. Proc. Roy. Soc. 180: 82. 1942.

length of the average freely orienting segment. There is no apparent necessity for abandoning equation (1) or for changing its form. Only β requires revision. Since it depends on the inverse first power of the length of the segment and on the inverse square root of the number of segments, β should decrease somewhat with a decrease in flexibility of the chain. The average displacement lengths will increase approximately as the square root of the average equivalent segment length. This segment length remains a rather indefinite quantity.

Bresler and Frenkel[6] have attempted a quantitative treatment of the effects of hindered rotation on chain configuration. They assume a potential given by equation (7) with $n = 1$, which is sufficiently great to limit the average rotation angle $\bar{\theta}$ to moderately small values, i.e., $U_0 \gg RT$. They derived an expression that may be written

$$\bar{r^2}/l^2 = 3Z(1 - \cos\omega)^2/(1 + \cos\omega)(1 - \cos\bar{\theta}) \tag{8}$$

and showed that the value of $\bar{\theta}$ is related to the potential U_0 according to

$$\cos\bar{\theta} = L(U_0/2RT), \tag{9}$$

where L is the Langevin function. Substituting equation (9) in (8) and recalling that $\cos\omega \cong -1/3$, one obtains

$$\bar{r^2} = 8l^2Z/[1 - L(U_0/2RT)]. \tag{10}$$

TABLE 1

U_0* (cal./mole)	U_0 (cal./mole)	$\left(\dfrac{\bar{r^2}\text{ (hindered rotation)}}{\bar{r^2}\text{ (free rotation)}}\right)^{1/2}$
9000	1000	5.48
3600	400	3.49
1800	200	2.67

The ratio of r calculated for hindered rotation using equation (10) to r for free rotation using equation (5) is shown in TABLE 1 for several values of the hindrance potential U_0. In drawing comparisons between these potentials and those for simple molecules such as ethane (3600 cal.),[7] propylene and acetone (600–2000 cal.),[8] it must be borne in mind that Bresler and Frenkel used a potential function with a single minimum, i.e., $n = 1$ in equation (7). These simpler molecules possess three potential minima ($n = 3$). For the small displacements from $\theta = 0$ assumed

[6] Bresler, S. E., & Frenkel, J. I. Acta Physicochimica USSR 11: 485. 1939.
[7] Kistiakowsky, G. B., Lacher, J. R., & Ransom, W. W. Jour. Chem. Phys. 6: 900. 1938.
[8] Schumann, S. C., & Aston, J. G. Jour. Chem. Phys. 6: 485. 1938, Wilson, E. B., Jr., & Wells, A. J. Jour. Chem. Phys. 9: 319. 1941, Telfair, D., & Pielemeier, W. H. Jour. Chem. Phys. 9: 571. 1941.

by Bresler and Frenkel in the development of their theory, the shape of the potential curve near its minimum is approximately the same regardless of whether $n = 1$ or 3. But in order to make the two curves coincide in the vicinity of $\theta = 0$ it is necessary to let U_0^* for $n = 3$ equal $9\ U_0$ for $n = 1$. Both U_0 and U_0^* values are included in TABLE 1. In the case of natural rubber, the hindrance potential should be of the order of 1000 cal. or less, in analogy with propylene. The corresponding $U_0^* = 9000$ cal. leads to a considerable increase in r over the free rotation value. However, if other minima, possibly a few hundred calories above the lowest, were taken into account, this length would be appreciably decreased.

We conclude, therefore, that hindrance to rotation about bonds of the chain will generally increase the mean displacement length of the chain over that which would prevail for free rotation. However, this increase will not be manyfold. If the chains are very long the mean displacement length will remain only a small fraction of the chain contour length. The expansion of the configuration with increasing hindrance to free rotation should exert a noticeable effect on the viscosities of dilute solutions of the polymer; the more extended the configuration the greater will be the viscosity of the solution for a given chain contour length at the same concentration.[3, 9]

THE MODULUS OF ELASTICITY

The fact that the form of the distribution function (1) for very long chains does not depend on hindrance to rotation about chain bonds is of the utmost importance. The parameter β is dependent upon chain flexibility, but in the equations for elastic constants derived from (1), β does not appear.

Guth and Mark[2] showed, as a consequence of equation (1), that the modulus of elasticity of a vulcanized rubber should be proportional to the absolute temperature and to the number of chains, i.e., to the number of cross linkages in the network structure.

Kuhn[3] derived a more explicit expression for the modulus of elasticity E from equation (1). In the undeformed state Kuhn considered that the rubber "molecules" conform to the distribution of chain displacement lengths given by equations (1), or (1′) and shown in FIGURE 1. Deformation produces a transformation of the distribution of chain displacement lengths to a new, less probable, distribution. Taking z to be the direction of the elongation, Kuhn assumed that the z-component of the

[9] Huggins, M. L. Jour Phys. Chem **42**: 911 1938, **43**: 439 1939, Jour. Applied Phys **10**, 700. 1939.

vector connecting the chain ends would be increased by a factor α, where α is the ratio of the final to the initial lengths of the sample. The x- and y-components were assumed to be decreased by the factor $1/\alpha^{1/2}$, since there is no change in volume on stretching rubber, except at high elongations. This assumption, which is basic to Kuhn's derivation, defines the new distribution as a function of the degree of deformation. From a consideration of the ratio of the probabilities of the deformed and the initial distributions, there is obtained for the entropy change on stretching

$$\Delta S = -\frac{3}{2} k\nu(\alpha - 1)^2, \tag{11}$$

where ν is the number of "molecules" in the sample and k is the Boltzmann constant. Since the change in heat content on moderately stretching rubber is small,[10] the free energy change on stretching is

$$\Delta F \cong \frac{3}{2} kT\nu(\alpha - 1)^2. \tag{12}$$

Letting f represent the force of retraction at length L, E the Young's modulus of elasticity, A the cross sectional area, and V the volume of the sample,

$$f = (\partial \Delta F/\partial L)_T = (1/L_0)(\partial \Delta F/\partial \alpha)_T = 3kT\nu(\alpha - 1)/L_0;$$

$$E_0 = (1/A_0)(\partial f/\partial \alpha) = 3kT\nu/A_0L_0 = 3kT(\nu/V) = 3RT\rho/M_c \tag{13}$$

The subscript "zero" refers to the undeformed state ($\alpha = 1$). The quantity ν/V is the number of chains per unit volume of rubber; ρ is the density of the rubber, and M_c is the molecular weight of one chain.

In Kuhn's earlier work there seems to be no clear conception of the nature of the entity which he termed the rubber "molecule." In subsequent papers he has recognized that the element of the network structure of a vulcanized rubber is the chain between two cross linkages. In his original paper on rubber elasticity he apparently considered such a chain to be a free and independent unit. Actually, its end points are constrained by their connections with the rest of the network. In considering the "molecule" (i.e., chain) to be an independent unit (which, if it were, would display no permanent elastic properties since it would be free to relax through internal reorganization), Kuhn sought to take into account lateral deformations of the molecule as well as changes in its length. In this way he arrived at a coefficient of 7/2 instead of 3/2 in

[10] Meyer, K. H., & Ferri, C. Helv. Chim. Acta 18: 570. 1935; Anthony, R. L., Caston, R. H., & Guth, E. Jour Phys Chem. 46: 826. 1942, Treloar, L. R. G. Trans Faraday Soc. 38: 293. 1942.

equation (11); in his expression for the elastic modulus corresponding to equation (13), Kuhn obtained a coefficient of 7.

Wall [11] has recently simplified the mathematical procedure for deriving the entropy change on stretching a vulcanized rubber. Starting with the distribution of lengths given by equation (1) and making the same assumption Kuhn employed regarding simple transformation of the x, y and z components of the r vectors, Wall obtains

$$\Delta S = -k\nu(\alpha^2 + 2/\alpha - 3)/2, \tag{11'}$$

which reduces to equation (11) for small deformations ($\alpha \leqq 1$). He has recognized from the nature of the network structure of a vulcanized rubber that only the ends of the chains are constrained to new (average) positions as the direct result of deformation. From equation (11′) there is obtained

$$E = kT(\nu/V)(1 + 2/\alpha^3) \tag{13'}$$

$$= RT\rho(1 + 2/\alpha^3)/M_c,$$

which reduces to equation (13) when $\alpha = 1$.[12, 13]

Bresler and Frenkel[6] attempted to treat the problem of the elastic reaction of long chains by employing a dubious analogy to diffusion theory. After correcting errors in numerical factors, introduced in the reduction of their equations, Bresler and Frenkel's treatment leads to

$$E = k^2T^2\ 2\nu l^2 U_0 \tag{14}$$

Contradicting experimental evidence that E increases with the first power of T, equation (14) contains the temperature raised to the second power. The derivation of equation (14) is based on a questionable analogy rather than on any straightforward application of statistical mechanics such as has been used by Kuhn and others. Furthermore, Bresler and Frenkel's equation (14) is limited to values of $U_0 \gg kT$.

Recently the authors[14] have carried out a statistical treatment of elastic properties of three-dimensional network structures which avoids entirely the necessity for the basic assumption of Kuhn and of Wall, namely, their assumption regarding the simple manner in which the chain-length distribution is transformed by macroscopic deformation of the sample. As an approximation to the actual structure we consider

[11] Wall, F. T. Jour. Chem. Physics 10: 132, 485. 1942.

[12] An equation similar to (11′) has been obtained by Guth, E., & James, H. M. Ind. Eng. Chem. 33: 624. 1941.

[13] Various authors (see, for example, Dostal, H. Monatshefte für Chemie 71: 144. 1938. Pelzer, H. Monatshefte für Chemie 71: 444. 1938) have obtained or discussed equations for the force of retraction which indicate that the rubber should shrink to zero length upon removal of the tension. This absurdity is the result of failure to take into account the random orientations of the initial r vectors with respect to the direction of stretch.

[14] To be published.

a network in which all chains (i.e., portions of the structure between two consecutive cross linkages) are of the same size, i.e., contain the same number Z of chain atoms. Each cross linkage represents a junction of four chains. The positions of the opposite ends of these four chains define a tetrahedron. The central junction in question may occupy any one of numerous positions, and the probability of any one position will depend on the distance to the four corners of the tetrahedron; that is, on the displacement lengths of the four chains leading to the respective corners.

We assume that the average restrictions to which each junction is subjected by the network in the undeformed state can be replaced by those which would result if the four nearest associated junctions were fixed at the corners of a regular tetrahedron. A deformation of the rubber produces a corresponding deformation in the elementary tetrahedron. The model is shown in FIGURE 2. The point P represents the central junction

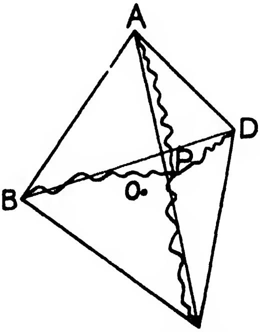


FIGURE 2. Tetrahedral model for an idealized cross-linked network structure.

in question. The four chains are represented schematically by wavy lines. The probability that the central junction lies in a volume element at P will depend on the product of four $W(r_i)$ functions given by equation (1), where the r_i assume the four values for the distances from P to A, B, C, and D. This probability decreases with the distance of P from the center O of the tetrahedron. Integrating throughout space we obtain a probability for the undeformed network structure. Upon repeating the same calculation for the deformed tetrahedron and comparing the probability so obtained with the above, the entropy of deformation is found to be

$$\Delta S = -k(\alpha^2 + 2/\alpha - 3)$$

per cross linkage. Multiplication of the above equation by the number $\nu/_2$ of cross linkages leads to equation (11′) for the entropy of deformation of the entire stock.

The equivalence of the expressions obtained by two such diverse procedures is reassuring. Equation (13′) affords a simple relationship between the fundamental elastic constant E and the most important parameter describing the network structure, namely, the concentration of cross linkages.

SWELLING PHENOMENA

When a high polymer is placed in contact with a suitable solvent, the polymer is observed to swell as the result of imbibition of the solvent. If the polymer molecules consist of essentially linear chain structures of finite size (though they may be very large) this swelling process is limited only by the amount of solvent available; the polymer is said to dissolve completely. If, however, the polymer possesses a continuous network structure, as in the case of a vulcanized rubber, the amount of liquid absorbed will reach a limit beyond which no further swelling will occur The greater the number of cross bonds in the gel the less it will swell when in contact with a given liquid. In the case of a "loose" gel structure, swelling may exceed a fiftyfold increase in volume; a tightly vulcanized rubber will swell only two- or threefold.

According to the most plausible explanation which has been advanced[15] for this phenomenon, absorption of solvent is due to osmotic "forces." The consequent expansion of the polymer network structure may be regarded as analogous to an elastic deformation. Equilibrium is attained when the osmotic forces are balanced by the elastic reaction of the network structure. Solvent in the gel then is in equilibrium with excess solvent. Specifically, the condition for swelling equilibrium requires that the partial molal free energy of the solvent in the gel shall equal the molal free energy of the unabsorbed solvent; for equilibrium with pure solvent, $\Delta \bar{F}_1 = 0$, where

$$\Delta \bar{F}_1 = \Delta \bar{F}_{m,1} + \Delta \bar{F}_{e,1}, \tag{15}$$

$\Delta \bar{F}_{m,1}$ being the osmotic and $\Delta \bar{F}_{e,1}$ the elastic contribution to the partial molal free energy. Alternatively,

$$\Delta \bar{F}_1 = \Delta \bar{H}_1 - T\Delta \bar{S}_{m,1} - T\Delta \bar{S}_{e,1}.$$

The term $T\Delta \bar{S}_{m,1}$ arises from the mixing of the chains with solvent.

[15] See for example, Frenkel, J. Acta Physicochimica USSR 9: 235. 1938, Gee, G. Trans. Faraday Soc. 38: 418. 1942.

Recently Huggins[16] and Flory[17] independently carried out statistical mechanical treatments of polymer–solvent mixtures which are in close agreement with the observed thermodynamic behavior of such systems.[18] On the basis of these investigations the partial molal entropy (due to mixing) of solvent in a mixture containing a volume fraction v_2 of polymer of very high molecular weight is

$$\begin{aligned} \Delta\bar{S}_{m,1} &= -R[\ln(1-v_2)+v_2] \\ &= Rv_2[v_2/2+v_2^2/3+\ldots]. \end{aligned} \tag{16}$$

The entropy change due to deformation (expansion) of the network can be readily calculated from the tetrahedral model previously discussed[14] in connection with the derivation of elastic moduli. Swelling merely magnifies the tetrahedron in proportion to the volume change. From the resulting change in the number of available states per junction we find

$$\Delta\bar{S}_{e,1} = -R\rho V_1 v_2^{1/3}/M_c, \tag{17}$$

where ρ is the density of the polymer and V_1 is the molar volume of the solvent; the ratio of the volume of the gel to that of the initial unswollen gel is $1/v_2$. Hence

$$\Delta F_1 = \Delta\bar{H}_1 + RT[\ln(1-v_2)+v_2+\rho V_1 v_2^{1/3}/M_c]. \tag{18}$$

From this equation the activity of the solvent can be calculated as a function of the concentration and of the degree of cross linking in the network.

From the extent of swelling at equilibrium ($\Delta\bar{F}_1 = 0$), the molecular weight per chain, which is inversely proportional to the concentration of cross linkages, can be calculated from the equation

$$M_c = -\rho V_1 v_2^{1/3}/[(\Delta\bar{H}_1/RT)+\ln(1-v_2)+v_2]. \tag{19}$$

If the heat of dilution can be neglected and the degree of swelling is large, i.e., if v_2 is small,

$$M_c \cong 2\rho V_1/v_2^{5/3}. \tag{19'}$$

Three-dimensional network structures such as occur in rubber vulcanizates evade most physicochemical experimental methods because of their insolubility in all solvents which do not destroy the structure. Equations (19) and (19′) provide a basis for evaluating polymeric ma-

[16] Huggins, M. L. Jour. Chem. Phys. 9: 440. 1941, Jour. Phys. Chem. 46: 151. 1942; Ann. N. Y. Acad. Sci. 43: 1. 1942.
[17] Flory, P. J. Jour. Chem. Phys. 9: 660. 1941; 10: 51. 1942.
[18] See for example, Gee, G., & Treloar, L. R. G. Trans. Faraday Soc. 38: 147. 1942.

terials possessing such structures. The equilibrium swelling method is made especially attractive by the facility with which such measurements can be carried out.[19]

SUMMARY

The statistical theory of the configuration of long polymer molecules has been examined in the light of more recent knowledge concerning hindrance of rotation about single valence carbon to carbon bonds. The *form* of the equation expressing the distribution of distances between the ends of the chains is unaffected by hindrance to rotation. In general, the average distance between chain ends is increased by hindrance to rotation. However, this increase will seldom be but a fraction of the difference between the chain "displacement length" and the contour length of the chain.

Various methods for calculating elastic moduli of polymeric materials possessing network structures (e.g., vulcanized rubbers) have been discussed. A new treatment developed by the writers leads to a force-elongation function identical with that derived by Wall by a revision of Kuhn's procedure.

Swelling of such materials by solvents can be treated as a combination of osmotic dilution opposed by the elastic reaction of the swollen network. Equations are derived for computing the concentration of cross linkages from the equilibrium degree of swelling. These equations should be particularly useful in the elucidation of the structures of gels, which evade most physicochemical methods of attack because of their inherent insolubility.

---

[19] **Scott, J. R.** Trans. Inst. Rubber Ind. 5: 3. 1929. **Blow, C. M., & Stamberger, P.** Rec. trav. chim. 48: 681. 1929. **Whitby, G. S., Evans, A. B. A., & Pasternack, D. S.** Trans. Faraday Soc. 38: 269. 1942.

THERMODYNAMIC PROPERTIES OF SOLUTIONS OF HIGH POLYMERS: THE EMPIRICAL CONSTANT IN THE ACTIVITY EQUATION*

By Maurice L. Huggins

From the Research Laboratories, Eastman Kodak Company, Rochester, N. Y.

INTRODUCTION

According to Raoult's law, the thermodynamic activity (a_1 or a_2) of each component of a solution is equal to its mole fraction (N_1 or N_2) in that solution. For solutions containing only small molecules, this law is known to hold well, if the heat of mixing of the components is negligible. For solutions of high polymers in small molecule solvents, however, this is not the case. Such solutions show very large deviations from Raoult's law, much larger than can be accounted for by the observed heats of mixing.

These deviations can be attributed to a large entropy of mixing effect, a result of a randomness of orientation of each segment of the solute molecule chain relative to the adjacent segments.

Using the methods of statistical mechanics, Flory[1] and the writer[2-4] (independently) have calculated the entropy of mixing of binary solutions of flexible long-chain molecules in small molecule solvents and have arrived at the following equations for the activities of the components.

$$\ln a_1 = \ln V_1 + \left(1 - \frac{\bar{V}_1}{\bar{V}_2}\right) V_2 + \mu_1 V_2^2, \tag{1}$$

$$\ln a_2 = \ln V_2 + \left(1 - \frac{\bar{V}_2}{\bar{V}_1}\right) V_1 + \frac{\bar{V}_2}{\bar{V}_1} \mu_1 V_1^2. \tag{2}$$

Here $\bar{V}_1$ and $\bar{V}_2$ are partial molal volumes. V_1 and V_2 are volume fractions, and μ_1 is a constant characteristic of the pair of components. This constant is the subject of the present paper.

These equations—either of which can be deduced from the other by the

* Communication No. 896 from the Kodak Research Laboratories.

[1] Flory, P. J. Jour. Chem. Phys. 9: 660. 1941; 10: 51. 1942.

[2] Huggins, M. L. Jour. Chem. Phys. 9: 440. 1941.

[3] Huggins, M. L. Jour. Phys. Chem. 46: 151. 1942.

[4] Huggins, M. L. Ann. N. Y. Acad. Sci. 43: 1. 1942.

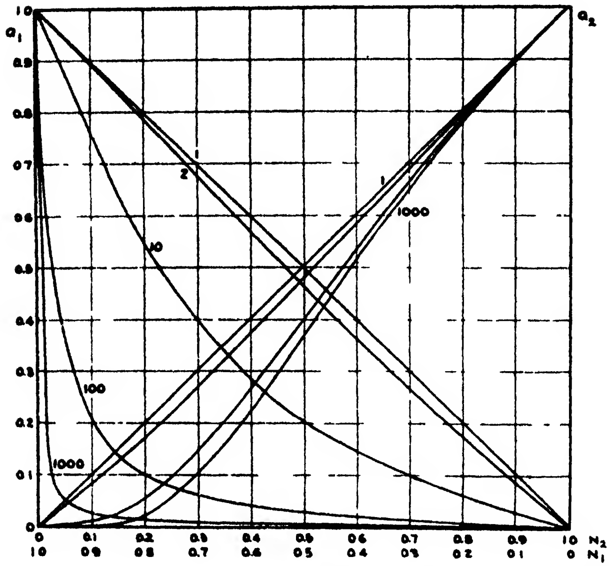


FIGURE 1. Variation with mole fraction of the activities of the components of a binary solution, for various values of $\bar{V}_2/\bar{V}_1$ (indicated by the numbers alongside the curves), with μ_1 equal to zero. The unnumbered curves of a_2 are for the partial molal volume ratio equal to 2 and 10 for this ratio equal to 100, the curve is practically identical with that for a ratio of 1000

Gibbs-Duhem-Margules relationship—have been tested[2-6] with experimental data from a wide variety of systems and have been found to agree satisfactorily with these data.

FIGURE 1 shows the dependence of the activity *vs.* mole fraction curves of the components on the partial molal volume ratio, $\bar{V}_2/\bar{V}_1$, according to the foregoing equations, for μ_1 equal to zero. FIGURES 2, 3 and 4 show activity *vs.* mole fraction and activity *vs.* volume fraction curves for other values of μ_1. For values of μ_1 greater than a critical value, given by the equation (1),

$$\mu_1(\text{crit.}) = \tfrac{1}{2}[1 + (\bar{V}_1/\bar{V}_2)^{1/2}]^2, \tag{3}$$

these curves are S-shaped, indicative of separation into two phases.

THE DEPENDENCE OF μ_1 ON THE HEAT OF MIXING

The magnitude of μ_1 depends on several factors, perhaps the most important being the heat of mixing. From thermodynamics, one can

[5] Huggins, M. L. Jour Amer. Chem. Soc. 64: 1712. 1942.
[6] Huggins, M. L. Ind. Eng Chem. 35: 216, 980 1943

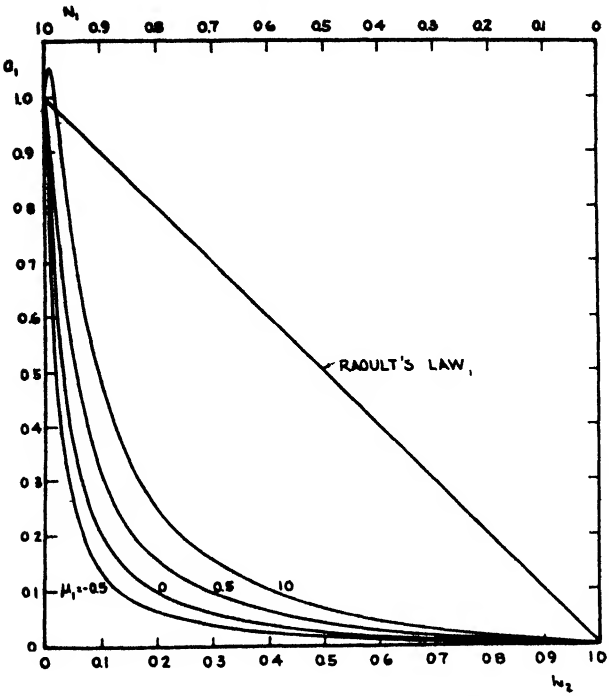


FIGURE 2. Activity (a_1) as a function of mole fraction for $\bar{V}_2/\bar{V}_1 = 100$, with certain values of μ_1.

show that the expression for ln a_1 must contain a term $\bar{L}_1/RT$, where $\bar{L}_1$ is the partial molal heat of mixing of component 1. If, following van Laar,[7] Scatchard,[8] and Hildebrand,[9] we assume $\bar{L}_1$ to be given approximately by the relation

$$\bar{L}_1 = K_{1,2}\bar{V}_1 V_2^2, \tag{4}$$

$K_{1,2}$ being a constant depending on the "internal pressures" in the pure liquid components, then one of the additive terms contributing to μ_1 has the magnitude $K_{1,2}\,\bar{V}_1/RT$.

Since equation (4) can hardly be expected to hold accurately at high concentrations of component 2, μ_1 should be expected to vary considerably with concentration at high values of V_2, in cases in which its

, J. J. van. Z. physik. Chem. A137: 421. 1928.
chard, G. Chem. Rev. 8: 321. 1931.
, J. H. Jour. Am. Chem. Soc. 57: 866. 1935; Chem. Rev. 18: 315. 1936; "Solubility s." Second edition. Reinhold Publishing Corp. New York. P. 73. 1936.

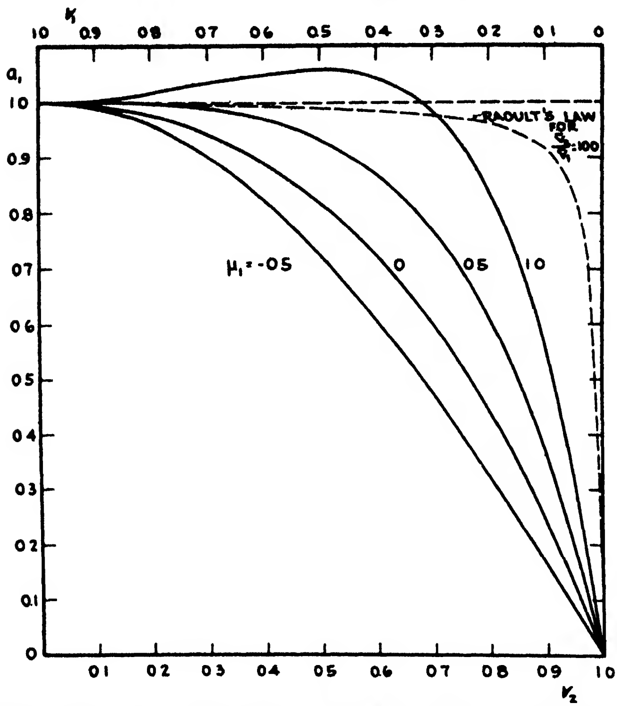


FIGURE 3. Activity (a_1) *vs* volume fraction, according to equation (1), for $\bar{V}_2/\bar{V}_1 = \infty$, with certain values of μ_1

magnitude is determined primarily by the heat of mixing. This may be the cause of the apparent variation of μ_1 with composition in certain solutions. (See below, under the heading "Variation of μ_1 with Concentration.")

Hildebrand[9] relates $K_{1,2}$ to the molal energies of vaporization of the pure liquid components by the equation

$$K_{1,2} = \left[\left(\frac{\Delta E_1}{V_1}\right)^{1/2} - \left(\frac{\Delta E_2}{V_2}\right)^{1/2}\right]^2 \tag{5}$$

If equations (4) and (5) were strictly true, the partial molal heat of mixing would obviously always have to be positive. Actually, heats of mixing are frequently negative. In such cases, one should be especially cautious about assuming equation (4) to hold, with $K_{1,2}$ constant, and therefore about assuming μ_1 to be invariant with concentration.

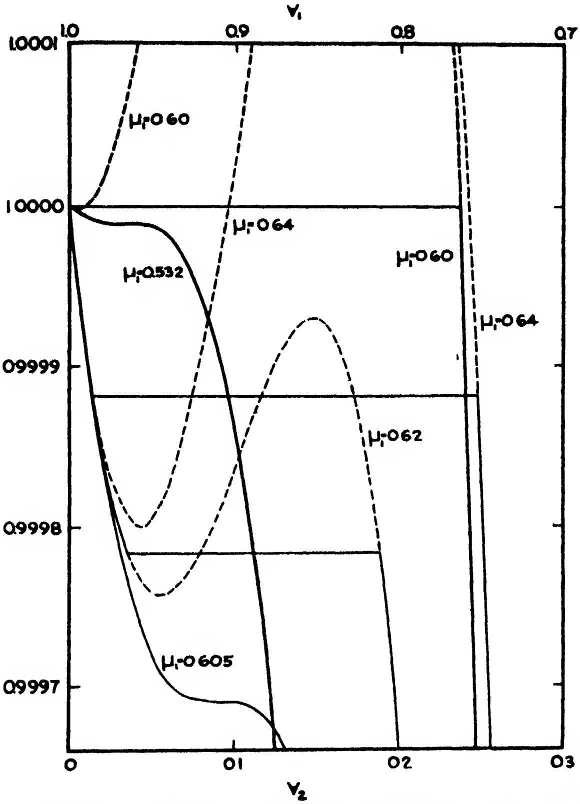


FIGURE 4. Activity (a_1) *vs* volume fraction, for $\bar{V}_2/\bar{V}_1$ equal to 100 (lighter lines) and 1000 (heavier lines), with certain values of μ_1. The non-realizable portions of the activity curves are shown dashed.

THE DEPENDENCE OF μ_1 ON OTHER FACTORS

Even with no heat of mixing, the statistical mechanical development leads to a finite value of μ_1, in the following way. The expression derived for the activity of component 1 is (see equation (36) of reference 4)

$$a_1 = V_1 \left(1 - \frac{2}{z'} V_2\right)^{-\frac{z'}{2}\left(1 - \frac{\mathrm{V}_1}{\mathrm{V}_2}\right)} \exp\left(\frac{L_1}{RT}\right), \tag{6}$$

in which z' is the effective average coordination number of the solvent molecules and solute submolecules. Taking the logarithm, this gives

$$\ln a_1 = \ln V_1 - \frac{z'}{2}\left(1 - \frac{\bar{V}_1}{\bar{V}_2}\right)\ln\left(1 - \frac{2}{z'}V_2\right) + \frac{L_1}{RT}. \tag{7}$$

Since

$$\ln\left(1 - \frac{2}{z'}V_2\right) = -\frac{2}{z'}V_2 - \frac{2}{z'^2}V_2^2 - \frac{8}{3z'^3}V_2^3 - \quad , \tag{8}$$

equation (7) may be put in the form

$$\ln a_1 = \ln V_1 + \left(1 - \frac{\bar{V}_1}{\bar{V}_2}\right)V_2 + \frac{1}{z'}\left(1 - \frac{\bar{V}_1}{\bar{V}_2}\right)V_2^2 + \frac{4}{3z'^2}\left(1 - \frac{\bar{V}_1}{\bar{V}_2}\right)V_2^3 + \quad + \frac{L_1}{RT} \tag{9}$$

For very large molecules, this reduces to

$$\ln a_1 = \ln V_1 + \left(1 - \frac{\bar{V}_1}{\bar{V}_2}\right)V_2 + \frac{1}{z'}V_2^2 + \frac{4}{3z'^2}V_2^3 + \quad + \frac{L_1}{RT} \tag{10}$$

For z', one may reasonably assume a value between 5 and 10. Therefore, at low concentrations (V_2 small), this theoretical treatment would indicate a value of μ_1 of 0.1 or 0.2, in addition to other contributions. The term in V_2^3 and higher terms are small, even at high concentrations, and may be neglected.

For hypothetical long-chain molecules assumed to be composed of segments, each consisting of a rigid string of submolecules (equal in size to the solvent molecules), the statistically derived equation for $\ln a_1$ is identical with equation (10), except for the replacement of z' by $\gamma z'$, where γ has a value in the neighborhood of 2 (for several submolecules per segment). Changing to this molecular model from one in which the chain is flexible at *every* connection between submolecules thus merely reduces the contribution to μ_1 discussed in the preceding paragraph to about half its already small value.

Consideration of the statistical development shows that the assumption that the several submolecules of each segment are in a straight line is of no importance; the important fact is that a certain proportion of the joints between submolecules is inflexible. *The shape of each segment is therefore unimportant; increasing its volume (relative to that of each solvent molecule) modifies equations (1) and (2) only to the extent of making μ_1 slightly (about 0.1, at most) less positive.*

In the chain-molecule model just discussed, zero flexibility was assumed at a fraction of the joints between submolecules, complete flexibility being assumed for the remainder. Qualitatively, at least,

this model should be equivalent—as regards entropy of mixing calculations—to another in which there is some flexibility, but only a limited amount, at each joint. One may conclude, therefore, that the activities of the components in a solution of chain molecules depend to only a slight extent on the flexibility of the chains. A large decrease in flexibility makes μ_1 slightly smaller in magnitude, if positive, or larger, if negative.

In deriving equations (1) and (2), the entropy of mixing was computed on the assumption of perfect randomness of mixing of the two molecular species. This assumption is certainly not correct if the density of attraction energy between like molecules is either larger or smaller than the density of attraction energy between unlike molecules. (This language is admittedly not precise, but it expresses the fundamental idea.) There will be a tendency either toward aggregation or toward solvation. The effect of this tendency on the heat of mixing is approximately taken care of, as indicated above, by the inclusion of a term $\left(\frac{K_{1,2}\bar{V}_1}{RT}\right)V_2^2$ in the expression for ln a_1. The effect on the entropy of mixing may be approximated by considering that y (the average number of alternative sites for each submolecule, except the first, in the chain polymer)[3, 4] changes as V_2 increases, in accordance with the relation

$$\frac{dy}{dV_2} = -k_y y V_2. \tag{11}$$

In general, one would expect k_y to have a positive sign for a positive heat of mixing ($\bar{L}_1$), the agglomerization of the chain molecules reducing the number of available sites for each succeeding submolecule. For a negative heat of mixing it would seem that the sign of k_y might be either negative or positive, depending in part on whether any decrease in flexibility of the chain molecules resulting from the attraction of solvent molecules to them (solvation) outweighs the increase in flexibility resulting from the fact that each chain submolecule has fewer other chain submolecules adjacent to it than for purely random mixing.

If equation (11) is assumed, the equation deduced for ln a_1 is identical with equation (1) except for the substitution of $\mu_1 - k_y$ for μ_1. At least for positive values of $\bar{L}_1$, this means that the effect of the unequal molecular attractions on the entropy of mixing produces a term contributing to μ_1 which is opposite in sign to the term resulting directly from the heat of mixing. Probably the entropy contribution is normally smaller than the heat contribution.

From the foregoing discussion one may predict that, if L_1/RTV_2^2 is positive and not too small, μ_1 will also be positive, but smaller in magnitude; if L_1/RTV_2^2 is zero or positive and small (perhaps less than 0.2 or 0.3), μ_1 will be positive and small (perhaps 0.1 to 0.3); and if L_1/RTV_2^2 is negative, μ_1 will have a smaller negative value.

THE DEPENDENCE OF μ_1 ON TEMPERATURE

A rise in temperature should make the departures from perfect mixing smaller. Both the heat of mixing contribution to μ_1 and the contribution to μ_1 of the entropy of mixing resulting from imperfect mixing of the components should therefore be smaller in magnitude the higher the temperature. For these contributions we may reasonably assume inverse proportionality with the absolute temperature. A rise in tempera-

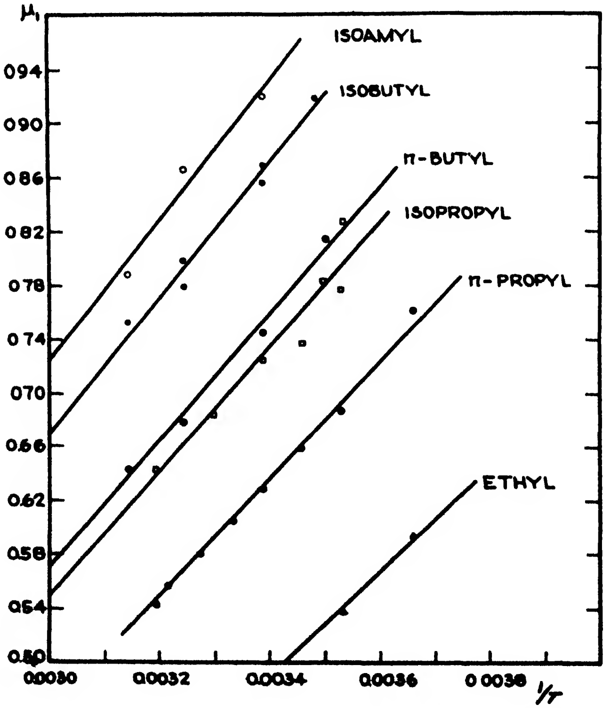


FIGURE 5. Variation of μ_1 with the reciprocal of the absolute temperature, for polystyrene swollen in alkyl laurates.

ture should also increase the flexibility of the chain molecules; this would make μ_1 slightly less positive. The other contributions to μ_1 just discussed are relatively small and probably not much affected by temperature changes.

Tentatively, then, we may assume the variation of μ_1 with temperature to be given by the relation

$$\mu_1 = \alpha_1 + \frac{\beta_1}{T}, \tag{12}$$

β_1 being related to V_1, the molal volume of the solvent, by the equation

$$\beta_1 = \gamma_1 + \delta_1 V_1, \tag{13}$$

at least when comparing solvents that are sufficiently similar in other respects. These relationships are in agreement with data[12] on polystyrene-alkyl laurate gels, as shown in FIGURES 5 and 6.

Values of α_1 and β_1 computed for the rubber-benzene system and for solutions of oleyl oleate in n-hexane and in cyclohexane are listed, with the polystyrene-alkyl laurate values, in TABLE 1.

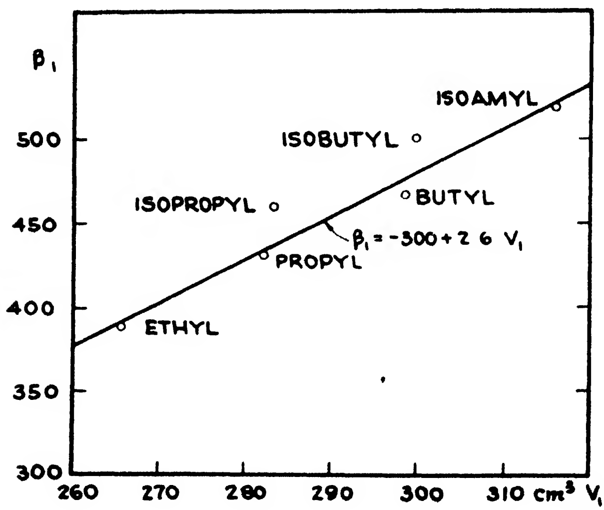


FIGURE 6. Dependence of β_1 on the molal volume of the small molecule component, for polystyrene-alkyl laurate gels.

TABLE 1
VALUES OF CERTAIN CONSTANTS

Components	μ_1 (25° C.)	α_1	β_1
Oleyl oleate–cyclohexane[10]	+ 0.37	− 0.92	385
Oleyl oleate–*n*-hexane[10]	+ 0.33	− 0.41	220
Rubber–benzene[11]	+ 0.44	+ 0.37	20
Polystyrene–ethyl laurate[12]	+ 0.47	− 0.83	388
Polystyrene–*n*-propyl laurate[12]	+ 0.62	− 0.83	431
Polystyrene–isopropyl laurate[12]	+ 0.71	− 0.83	460
Polystyrene–*n*-butyl laurate[12]	+ 0.74	− 0.83	467
Polystyrene–isobutyl laurate[12]	+ 0.85	− 0.83	500
Polystyrene–isoamyl laurate[12]	+ 0.91	− 0.83	518

VARIATION OF μ_1 WITH CONCENTRATION

Most of the experimental data available for computing μ_1 values for polymer solutions are for low concentrations of polymer. For practically all of these solutions, μ_1 is constant within the probable experimental error over the concentration range studied.

Even for solutions and gels of high polymer concentration, μ_1 is

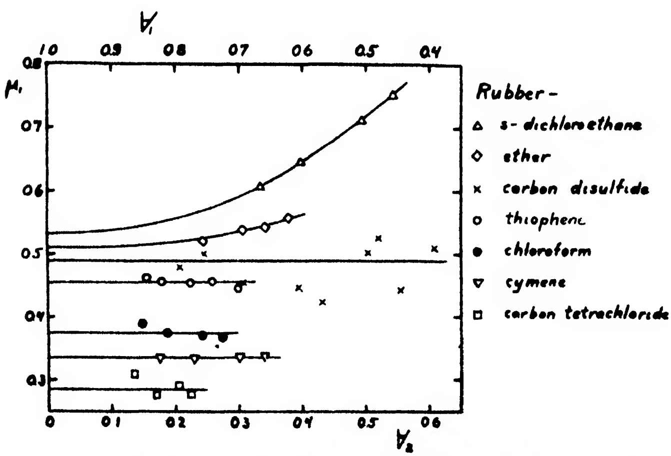


FIGURE 7 Dependence of μ_1 on concentration, for certain rubber-solvent systems. Solvent

Carbon tetrachloride[13]	Carbon disulfide[14]
Cymene[13]	Ether[13]
Chloroform[13]	*s*-Dichloroethane[13]
Thiophene[13]	

[10] Meyer, K. H., & Lühdemann, R. Helv. Chim. Acta. 18: 307. 1935.
[11] Gee, G., & Treloar, L. R. G. Trans. Faraday Soc. 38: 147. 1942. Dr. Gee, in a private communication, states that equation (20) of this paper should be corrected to read $\Delta h_0 = 2.0\left(\frac{w_r}{1.7 - 0.7 w_r}\right)^2$
[12] Brønsted, J. N., & Volqvartz, K. Trans. Faraday Soc. 35: 576. 1939.

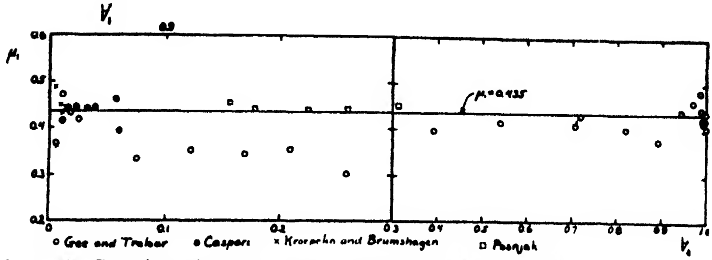


FIGURE 8. Dependence of μ_1 on concentration, for the rubber benzene system
Kroepelin and Brumshagen,[15] 40° C
Caspari,[16] 25° C
Posnjak,[13] 15° 20° C
Gee and Treloar,[11] 25° C

apparently quite independent of concentration in many cases (FIGURES 7 and 8). For the rubber-benzene systems, in fact, this constancy extends over the whole range from pure benzene to pure rubber.

Calculation of μ_1 from Posnjak's[13] swelling-pressure data on various rubber-solvent systems in the range between $V_2 = 0.15$ and $V_2 = 0.6$ shows it to depart significantly from constancy only in the two instances for which its value is largest (FIGURE 7). Although the apparent lack of constancy in these cases may be the result of experimental inaccuracies at high swelling pressures, it is just as likely that the deviations are real, for reasons discussed above.

VALUES OF μ_1 FOR DIFFERENT SYSTEMS

In TABLE 2 are collected values of μ_1 for a large number of different systems. Insofar as μ_1 is constant over the composition range of interest, the variation with concentration of the solvent and solute activities may be computed, using these μ_1 values, by means of equations (1) and (2).

TABLE 2

VALUES OF μ_1

Components	μ_1	Temperature (° C)
Rubber–benzene + 10 per cent ethanol[17]	0 26	25
Rubber–carbon tetrachloride[13]	28	15–20
Rubber–camphor[18]	29	180
Rubber–cymene[13]	.33	15 20
Rubber–cyclohexane[19]	33	6
Rubber–tetrachloroethane[13]	36	15–20
Rubber–chloroform[18]	37	15–20

[13] Posnjak, E. Koll.-Chem. Beih. 3: 417. 1912.
[18] Stamberger, P. Jour. Chem. Soc. 2318. 1929.

TABLE 2 (*Continued*)

VALUES OF μ_1

Components	μ_1	Temperature (° C.)
Rubber–cumene[13]	.38	15–20
Rubber–light petroleum[16]	.43	25
Rubber–*s*-dichloroethylene[13]	.43	15–20
Rubber–toluene[20, 21]	.43	27
Rubber–benzene[11, 13, 15, 16, 31]	.44	25
Rubber–chlorobenzene[22]	.44	7
Rubber–thiophene[13]	.45	15–20
Rubber–carbon disulfide[14]	.49	25
Rubber–amyl acetate[17]	.49	25
Rubber–benzene + 15 per cent methanol[17]	.50	25
Rubber–ether[13]	.51+	15–20
Rubber–*s*-dichloroethane[13]	.53+	15–20
Polyethylene oxide–water[18]	.45	27
Polystyrene–benzene[30]	.2	5
Polystyrene–toluene[28]	.44	27
Polystyrene–ethyl laurate[12, 31]	.47	25
Polystyrene–*n*-propyl laurate[12, 31]	.62	25
Polystyrene–*iso*propyl laurate[12, 31]	.71	25
Polystyrene–*n*-butyl laurate[12, 31]	.74	25
Polystyrene–*iso*butyl laurate[12, 31]	.85	25
Polystyrene–*iso*amyl laurate[12, 31]	.91	25
Polyvinyl chloride–tetrahydrofuran[32]	.14	27
Polyvinyl chloride–dioxane[32]	.52	27
Hydrogenated polyindene–benzene[30]	6	5
Copolymers of polyvinyl chloride and polyvinyl acetate–dioxane[32]	.4	27
Chlorinated polyvinyl chloride–dioxane[32]	.37	27
Gutta percha–carbon tetrachloride[20]	.28	27
Gutta percha–toluene[20, 23]	.36	27
Gutta percha–benzene[16]	.52	25
Balata–toluene[20]	.36	27
Hydrorubber–toluene[20]	.45	27
Cyclised rubber–toluene[20]	.46	27
Cellulose nitrate–cyclohexanone[24, 25]	.15	25
Cellulose nitrate–acetone[24, 26]	.19	20
Cellulose nitrate–acetone[24, 27]	.26	22
Cellulose nitrate–acetone[24, 28]	.30	27
Cellulose acetate–tetrachloroethane[24, 29]	− 1.8	24.4

15 **Kroepelin, H., & Brumshagen, W.** Ber. 61: 2441. 1928.
16 **Caspari, W. A.** Jour. Chem. Soc. 105: 2139. 1914.
17 **Gee, G.** Trans. Faraday Soc. 36: 1171. 1940.
18 **Gee, G.** Trans. Faraday Soc. 38: 109. 1942.
19 **Kemp, A. R., & Peters, H.** Ind. Eng. Chem. 33: 1263. 1941.
20 **Staudinger, H., & Fischer, K.** Jour. prakt. Chem. 157: 19. 1940.
21 **Meyer, K. H., Wolff, E., & Boissonnas, C. G.** Helv. Chim. Acta 23: 430. 1940.
22 **Fikentscher, H.** See **Meyer, K. H., & Mark, H.** Ber. 61: 1939. 1928.
23 **Wolff, E.** Helv. Chim. Acta 23: 439. 1940.
24 The μ_1 values characteristic of cellulose esters depend markedly on the degree of esterification.
25 **Boissonnas, C. G., & Meyer, K. H.** Helv. Chim. Acta 20: 783. 1937.
26 **Duclaux, J., & Wollman, E.** Compt. rend. 152: 1580. 1911.
27 **Dobry, A.** Jour. chim. phys. 32: 50. 1935.
28 **Schulz, G. V.** Z. physik. Chem. A176: 317. 1936.
29 **Haggar, O., & van der Wyk, A. J. A** Helv. Chim. Acta 23: 484. 1940.
30 **Kemp, A. R., & Peters, H.** Ind. Eng. Chem. 34: 1097. 1942.
31 See TABLE 1.
32 **Staudinger, H., & Schneiders, J.** Ann. 541: 151. 1939.

Most of the data on which the values of μ_1 in this table are based are for quite dilute solutions. In the few cases for which the data indicate a dependence of μ_1 on concentration, the value in the table is that obtained from the data at the lowest concentrations.

SUMMARY AND CONCLUSION

The importance of μ_1 results from the fact that the activities of the components of a binary solution (of polymers or otherwise) can be computed quite accurately over a large range of composition, once the value of μ_1 for the system in question is known. In this paper the available empirical information about values of μ_1 and about the variation of μ_1 with temperature and with concentration has been collected and the theoretical dependence of μ_1 on these and other factors has been discussed.

Discussion of the dependence of μ_1 on the specific chemical nature of the components will be left for another paper. Studies of this dependence can be made with the aid of data on solutions containing only small molecules as well as with data from polymer solutions, since, to a pretty good approximation at least, μ_1 is independent of the molecular weights and volumes of the components.

In conclusion, the writer is pleased to acknowledge his indebtedness to Mrs. Dorothy Owen Davis for help with the calculations required for the preparation of this paper.

Annals of The New York Academy of Sciences
Volume XLIV, Art. 5. Pages 445–538
December 14, 1943

SULFONAMIDES *

By

Colin M. MacLeod, Paul H. Bell, Henry Irving Kohn, J. S. Lockwood, Richard O. Roblin, Jr., James A. Shannon, and H. B. van Dyke

CONTENTS

* This series of papers is the result of a conference on Sulfonamides held by the Section of Physics and Chemistry of The New York Academy of Sciences April 16 and 17, 1943.

Publication made possible through a grant from the Conference Revolving Fund

INTRODUCTION TO THE CONFERENCE ON SULFONAMIDES

By Colin M. MacLeod

New York University College of Medicine, New York, N. Y.

In the short space of time since their introduction, the sulfonamide drugs have been the cause of a radical change in the approach to chemotherapy, not only as it has affected their own use, but that of unrelated compounds also. There are several reasons why this has occurred, among which may be mentioned the relatively simple chemical structure of the parent compound, sulfanilamide, the ease with which the activity of this group of compounds can be tested experimentally both *in vitro* and *in vivo*, and the emphasis early in the course of their study on the importance of understanding their disposition in the body. As a consequence, rapid advances have been made in the development of more potent derivatives, in modifications leading to a favorable distribution in the body and at the same time reducing the toxicity, and also in the knowledge of the mechanism of their antibacterial action

The paper of Dr. Roblin and Dr. Bell, which is here presented in abstract form, is concerned with the relation of certain physicochemical properties of sulfonamides to their antibacterial action In Dr. Shannon's communication, an account is given of the relation of the physicochemical structure of a series of compounds to their physiological disposition. From the observations recorded in these papers it is evident that considerable progress has been made in relating structure to function. However. as pointed out by Dr. Van Dyke, the information available on the toxic action of sulfonamides does not permit generalizations enabling one to predict the toxicity of a potentially useful compound in terms of its structure or some other physicochemical characteristic.

The important subject of primary and secondary inhibitors of sulfonamide activity is discussed by Dr. Kohn. In the final paper of the series, Dr. Lockwood deals with the therapeutic action of the drugs. particularly as it is related to the nature of the infectious process.

THE RELATION OF STRUCTURE TO ACTIVITY OF SULFANILAMIDE TYPE COMPOUNDS *

By

RICHARD O. ROBLIN, JR., AND PAUL H. BELL

From the
Stamford Laboratories, American Cyanamid Company, Stamford, Connecticut

GENERAL CONSIDERATIONS

For some time we have been interested in the relationship which it seems should exist between the molecular structure of sulfanilamide type compounds and their chemotherapeutic activity. Our approach to this problem has been through an attempt to utilize some fundamental physical property which is related both to structure and activity. Several factors must be considered before such an approach to this problem can be made. First, a reasonably accurate and reproducible method for the evaluation of chemotherapeutic activity is essential. Second, the mode of action of the compounds is of great importance. Finally, a fundamental physical property related to both structure and activity must be found. If the bacteriostatic action of sulfanilamide type compounds is due to a competition with the required *p*-aminobenzoic acid, then the more closely the competitor compound resembles this acid, the greater should be its blocking or bacteriostatic effect. Dissociation constants were selected because this readily measurable property furnishes a means of comparing the relative positive or negative character of various chemical groups.

The present theory is based on the experimental observation that the acid dissociation constants of N^1-substituted sulfanilamide derivatives are related to their chemotherapeutic activity.[1] If these constants are plotted against the activity, a smooth curve is obtained, which passes through a maximum as the acid strength increases Both the acid and base constants of more than one hundred sulfanilamide type compounds have been determined. The base constants all fall within narrow limits, whereas the acid constants vary over a wide range ($<10^{-11}$ to 10^{-3}). We believe that the correlation between acid constants and activity is directly associated with the relative negative character of the

* A detailed discussion of the relation of structure to activity in sulfanilamide-type compounds and the implications of the theory are given in the paper by P H Bell and R O Roblin, Jr, in the Journal of the American Chemical Society, volume 64 page 2905, 1942 At the Conference of Sulfonamides in April, 1943, the data presented in this paper were discussed in detail by the authors. An abstract of the material is presented in this article (Editor)

[1] Bell, P. H., & Roblin, R. O., Jr. Jour. Am. Chem. Soc 64: 2905 1942.

SO_2 group. Geometrically it can be shown that the SO_2 group of the sulfonamides and the CO_2 ion of *p*-aminobenzoic acid are very similar. Since the CO_2 ion is a strong negative group, it seems logical that the more negative the SO_2 group, the more closely it will resemble the CO_2 ion. The theory may then be stated as follows: *the more negative the SO_2 group of an N^1-substituted sulfanilamide derivative, the more bacteriostatic the compound will be.*

The acid constants can be shown to furnish an indirect measure of the negative character of the SO_2 group. Naturally, the proportion of ions to molecules will vary with acid strength in a buffered medium. Since it appears that the SO_2 group in the ionized form is more negative, it follows that the ionic form of any sulfonamide should be more active than the molecular form. However, it also appears that the negative character of the SO_2 group decreases with increasing acidity. Therefore, the more acidic the sulfonamide, the less negative the SO_2 group of the ionic and molecular forms, and the less the bacteriostatic activity of either form. Up to a certain point this factor will be more than compensated for by the increasing proportion of the highly active ionic form. Consequently, as acid strength increases, bacteriostatic activity should increase until the change in the proportion of ions to molecules is less significant than the decreasing activity of both forms. The net result of these interrelated factors may be used to explain the appearance of a maximum in the curve relating acid strength and bacteriostatic activity. A minimum is also predicted in this curve when the sulfonamides are so weakly acidic that the ionic form can be neglected.

Since the acid constants are related to both the structure of an N^1-substituted compound and its activity, an indirect correlation between structure and chemotherapeutic activity is established. Knowing something about the relative positive or negative character of the N^1-substituent, it is possible to predict with considerable accuracy the activity of a new sulfanilamide derivative of this type. A detailed discussion of the relation of structure to activity, and the implications of the theory, are given by Bell and Roblin.[1]

CHEMOTHERAPEUTIC ACTIVITY

As pointed out in published data,[1] the basic strengths of the *p*-amino groups of *p*-aminobenzoic acid and the sulfanilamides are of the same order of magnitude [$(0.5–2.3) \times 10^{-12}$]. In view of the fact that these variations were small and could not be correlated with activity, they were considered not be be important in determining activity, as long as they fell in this range of basic strength. It then was possible to quan-

titatively treat bacterial activity as a function of the acidity of the sulfonamide group.

As pointed out above,[1] the problem is apparently one of obtaining a proper balance between the acid strengthening effect of the R group and the formal ionic charge on the sulfonamide nitrogen to give the maximum over-all negative character to the SO_2 group. Branch and Calvin[2] have shown that the dissociation constant of an organic acid can be predicted quantitatively by an equation of the type

$$\log K = \log K_a + \Sigma I_R \alpha^i \qquad (1)$$

where K_a is the acid constant of the parent acid, I_R the inductive constants for each atom or group other than hydrogen, α the fraction that reduces the inductive effect for the transmission across each bond, and i the number of bonds through which the effect must be transmitted. I_R multiplied by 2.3 RT then becomes a potential and has the units of free energy.

Assuming that bacteriostatic activity is proportional to the potential of the SO_2 group, then, when the total activity is due almost entirely to the highly active ions (pK_a 2-11),

$$2.303\, RT \log (k/xC_R) = 2.303\, RT[\alpha(12.3 - I_R\alpha) - I_R\alpha^2]. \qquad (2)$$

where C_R = minimum molar concentration of a sulfonamide required to exhibit a given bacteriostatic activity, attributing all activity to the ions; x = fraction of the total concentration of the compound in ionic form; k = proportionality constant (determined experimentally) to adjust the potential energy of the SO_2 to experimental conditions, I_R; α, as defined for equation (1), $(12.3 - I_R\alpha)$ = inductive effect of the ionic charge reduced by the effect of I_R on it, $I_R\alpha^2$ = inductive effect on R on SO_2 directly.

As a first approximation, resonance and polarization effects were neglected, and α was taken as 1/2.8 (the value of Branch and Calvin[2] for a covalent bond); then

$$\log \frac{k}{xC_R} = 4.04 - 0.255\, I_R. \qquad (3)$$

Using Branch and Calvin's inductive constants for various radicals, it was found that I_R was a linear function of the pK_a values of the corresponding sulfanilamides, and given by the following equation

$$I_R = -\,1.33\, pK_a + 13.88; \qquad (4)$$

also, for any acid at pH 7 (where the *in vitro* tests were made)

$$pK_a = 7 - \log x/(1 - x). \qquad (5)$$

[2] Branch, G. E., & Calvin, M. "The Theory of Organic Chemistry." Prentice-Hall, Inc. New York, N. Y. 1941.

Substituting equations (4) and (5) in equation (3), we have

$$\log 1/C_R + \log k = 3.23 + 0.661 \log x + 0.339 \log (1 - x). \quad (6)$$

At conditions of maximum activity ($\log 1/C_R$ = max.)

$$\frac{d \log 1/C_R}{dx} = 0 = \frac{0.661}{x} - \frac{0.339}{1 - x};$$

$x = 0.661$ at maximum activity.

This corresponds to a sulfanilamide derivative with a pK_a of 6.7 as calculated by equation (5) and agrees very well with the experimentally observed maximum. This value of pK_a for maximum activity is independent of the *in vitro* tests and depends only on the inductive effects of the R groups.

Using the experimental maximum activity, $\log 1/C_R = 6.1 \pm 0.3$, k may be evaluated by substituting $\log 1/C_R = 6.1$ and $x = 0.661$ in equation (6) and solving for k.

The final ion activity equation is then

$$\log 0.001/C_R = 3.23 + 0.661 \log x + 0.339 \log (1 - x). \quad (7)$$

A similar equation may be derived for the activity of the un-ionized form

$$\log \frac{0.001}{(1 - x)C_R} = - 1.3\alpha - 2I_R\alpha^2, \quad (8)$$

from which it can be shown that its contribution toward activity will be small except for very large pK_a values.

Combining equations (8) and (3) we may show that

$$\log \frac{(1 - x)C_R \text{ (un-ionized)}}{xC_R \text{ (ionized)}} = 4.85, \quad (9)$$

which means that any ion is approximately $10^{4.85}$ times more active than the corresponding molecule. Now, if C_R (un-ionized) = C_R (ionized)

$$\log (1 - x)/x = 4.85$$

and pK_a must be 11.85 to fulfill this condition. At this pK_a, log 2.0 should be added to $\log 1/C_R$ from equation (6). In this way, the ion activity curve was corrected for the molecular activity at pK_a's greater than 10, and the total activity curve drawn.

At pH 7 the experimental and theoretical curves are in very good agreement from pK_a 10–5, but at lower values of pK_a the compounds are less active than predicted. The limitations in the development of the theory form an important factor contributing to this deviation. As pointed out by Branch and Calvin, an exact equation of the type of equation (1) should contain summation terms for the polarization and

resonance, as well as the inductive effect. While polarization is probably a function of pK_a, the relationship cannot be readily established experimentally. On the other hand, resonance is probably dependent on the specific character of the R group, and no general relationship between resonance and inductive effect can be established. In equation (2) polarization and resonance have been neglected, and therefore α is not equal to 1/2.8 (the simple covalent bond value) unless these effects are small. Where R is not too electronegative (weaker acids) the value used for α probably is a good approximation. However, as R becomes strongly electronegative, the SO_2-N-R bonds should become more ionic, and α should be greater than 1/2.8. Such an increase in α would give better agreement between the theoretical and experimental curves in the low pK_a range.

These arguments have assumed that the sulfonamides could freely reach the site of action and that the concentrations at this point are equal to that of the bulk of the medium. If the competition between p-aminobenzoic acid and the sulfanilamides takes place within the bacterium, then the permeability of the bacterium to the sulfanilamide becomes important. Cowles[3] believes that the bacterial wall is impermeable to the ionic form of the sulfanilamide and that the activities of the compounds are determined only by the acid strength that regulates the quantity of un-ionized forms which may diffuse into the cell. A large amount of evidence has appeared in the literature concerning low ionic diffusion through living membranes.[4] In this connection the recent electrophoretic work of Bradbury and Jordan,[5] concerning the effect of sulfanilamides on the mobilities of *E.coli*, may throw some light on the diffusion into bacteria. For sulfanilamides that are very highly ionized, this ion-blocking effect may be important and contributes toward the deviation of the experimental and theoretical curves, for compounds with pK_a values greater than 5

Cowles' theory does not satisfactorily explain the results of Fox and Rose[6] or Schmelkes and associates.[7] Fox and Rose calculated the p-aminobenzoic acid to sulfanilamide ion concentration ratios, at the point of reversal of inhibition, for four sulfanilamides of varying degrees of acidity. These same ratios may be calculated from the data of Schmelkes and associates, from which it may be shown that

[3] Cowles, P. B. Yale Jour. Biol. & Med. 14: 599. 1942
[4] "Cold Spring Harbor Symposia on Quantitative Biology." 8. 1940.
[5] Bradbury, F. R., & Jordan, D. O. Biochem. Jour. 36: 287. 1942
[6] Fox, C. L., & Rose, H. M. Proc. Soc. Exp. Biol. & Med. 50: 142. 1942.
[7] Schmelkes, F. C., Wyss, O., Marks, H. C., Ludwig, B. J., & Stranskov, F. B. Proc. Soc. Exp. Biol. & Med. 50: 145. 1942.

$$\frac{[p\text{-aminobenzoic acid}]}{[\text{sulfanilamide ion}]} \text{ decreases as } pK_a \text{ decreases.}$$

We have extended these measurements to more acid compounds and have found the same results. This observation is in agreement with the proposed theory; that is, that the ionic form of the more acid sulfanilamides is less active.

The experimental measurements are presented in more detail in reference 1. The effects of polarity, ionic character, and geometry of N^1 substituents on the activities of sulfanilamides are also discussed, from the standpoint of ability to compete with *p*-aminobenzoic acid and absorption into the blood stream and bacterial cell.

THE RELATIONSHIP BETWEEN CHEMICAL STRUCTURE AND PHYSIOLOGICAL DISPOSITION OF A SERIES OF SUBSTANCES ALLIED TO SULFANILAMIDE *

By James A. Shannon

From the
Department of Medicine,
New York University College of Medicine
and
The Research Service,
Third Medical Division,
Goldwater Memorial Hospital,
Welfare Island, New York

The activity of a chemotherapeutic agent results from its ability to participate in or interfere with some phase of biological activity. The specific as well as the over-all activity of the agent, however, is conditioned to a considerable extent by those factors that determine its ability to reach the specific site of action, the concentration it achieves at that site, and the length of time an effective concentration is maintained. It follows, therefore, that a change in chemical structure that is accompanied by a change in physiological disposition will have important consequences to the over-all chemotherapeutic activity of the resulting substances. A change in the chemical structure of an agent that produces a desirable effect or removes an undesirable effect from the standpoints of absorption, distribution, or excretion may also be expected to have effects on the specific activity of the substance. Consequently, a chemotherapeutic agent (in the present series, an antibacterial agent) must have two completely different sets of characteristics for it to be generally useful. The first of these relates to the host, the second to the bacterium. The following discussion is limited to a consideration of some of the characteristics of compounds allied to sulfanilamide that relate exclusively to the host

* This investigation has been aided by a grant from the John and Mary R. Markle Foundation. The results of these investigations have been prepared for publication in detail. (Fisher, S. H., Troast, L., Waterhouse, A., and Shannon, J. A. Jour. Pharm. & Exp. Therap. In press.) The sulfonamides used were generously supplied, as follows: N^1-amino sulfanilamide, by Dr. E. K. Marshall, Jr., Department of Pharmacology, The Johns Hopkins University College of Medicine, N^1-ethanol and N^1-hydroxy sulfanilamide by Eli Lilly Company, N^1-methyl and N^1-ethyl sulfanilamide by Winthrop Chemical Company, N^1-acetyl sulfanilamide by Schering Corporation, N^4-sulfanilyl sulfanilamide by Alba Pharmaceutical Company, N^1-4-amino phenyl sulfanilamide by Dr. G. L. Webster, University of Illinois College of Pharmacy, N^1-phenyl sulfanilamide by The Upjohn Company, sulfamethylthiazole by The Squibb Institute for Medical Research; and the remainder, from The American Cyanamid Company.

Information has been obtained on the distribution and excretion of some 30 derivatives of sulfanilamide and closely allied substances The distribution studies define the distribution of each compound in the body as a whole and in specific organs, and the facility with which each passes the blood-brain barrier. The excretion studies were performed in a manner that permits an estimation of the extent to which a compound participates in processes of renal tubular excretion or is reabsorbed in the renal tubule. The extent to which either of these processes proceeds, together with the rate of filtration, determines the over-all excretion rate of a compound. Observations have also been made on the ability of each compound to form reversible combinations with the nondiffusable constituents of plasma, presumably plasma albumen

EXPERIMENTAL TECHNIQUES

The presentation will be somewhat simplified by a description of experimental techniques in relation to the physiological disposition of a single substance Most of the distribution studies were performed on cats However, a sufficient number of observations were made on the dog to indicate that the data may be more generally applied The animals were lightly anesthetized with nembutal, an abdominal incision made, the renal pedicles ligated, and the wound closed. A known amount of the substance to be examined was then injected intravenously Blood samples were drawn at one and two hours or at the termination of the experiment, at which time a sample of cerebrospinal fluid was withdrawn from the cisterna and samples were taken of brain, lung, liver, pancreas, muscle and sciatic nerve.

The blood samples were centrifuged immediately and the separated cells and plasma refrigerated until analysis. They were then diluted and the protein removed by precipitation with trichloracetic acid at high dilution The tissue samples were ground with washed silica or in a tissue homogenizer. Sufficient trichloracetic acid was added to precipitate the proteins completely. The resulting mixture was extracted with gentle agitation for 30 minutes and the protein-free filtrate separated by centrifugation and filtration. The analysis of sulfonamide concentration was by a commonly accepted method.†[1]

The data contained in TABLE 1 are derived from such an experiment with sulfanilamide, but, as in the tables that follow, the presentation is

† All compounds studied are recoverable quantitatively when added to plasma providing plasma protein precipitation is at a high dilution Recovery from muscle and brain is incomplete in isolated cases. However, there is no instance where recovery is sufficiently incomplete to warrant extensive comment here (see reference in footnote, page 455)

[1] Bratton, A. C., & Marshall, E. K., Jr. Jour. Biol. Chem. 128: 537. 1939

TABLE 1

THE DISTRIBUTION OF SULFANILAMIDE IN THE CAT

9:25—Nembutal Kidneys ligated	11:00—B_1 plasma—3.10 mg. %
10:00—27.8 mg./K. sulfanilamide	12:00—B_2 plasma—2.88 mg. %

TISSUE SAMPLES

Plasma concentration	Ratio: concentration in tissues / concentration in plasma/0.93			Volume of distribution
mg./ml.	Muscle	CSF	Brain	% Body Weight
2.88	1.07	0.73	0.70	96.5

limited to the data from a few representative tissues. The experiment is quite typical in that a distribution equilibrium appears to have been approximated at one hour, because the plasma concentration of sulfanilamide at that time is much the same as at the end of the two-hour interval. It is apparent from the data that sulfanilamide not only diffuses readily into all tissues[2, 3] but it is localized in the cells of many.[4] This conclusion stems from finding that the volume of distribution of the sulfanilamide is in excess of the water content of the body and that the ratio of the sulfanilamide concentration in tissue to its concentration in plasma water is greater than is true for water itself. Were sulfanilamide freely diffusible throughout the body and not localized in any tissue, its volume of distribution would be in the order of magnitude of 65 per cent of the body weight and the tissue/plasma distribution ratios would be considerably lower than those observed. On the other hand, if the sulfanilamide were largely retained in an extracellular position, it would have a volume of distribution of the order of magnitude of 25 to 30 per cent and its distribution ratios would be of the order of magnitude as those of sodium or chloride. In the case of muscle, this would be in the range 0.11 to 0.12.[5] The presence of sulfanilamide in a relatively high concentration in the cerebrospinal fluid is in keeping with the observations of others[2] and indicates that the compound penetrates the blood-brain barrier quite freely. The absence of a concentration identical to that of plasma is in part due to the factor of plasma binding, in this case about 10 per cent, and probably in part to a fairly

[2] **Marshall, E. K., Jr., Emerson, K., & Cutting, W. C.** Jour. Pharm. and Exp. Therap. 61: 196. 1937.
[3] **Painter, E. E.** Am. Jour. Physiol. 129: 744. 1940.
[4] **Waterhouse, A., & Shannon, J. A.** Proc. Soc. Exp. Biol. and Med. 40: 481. 1939.
[5] **Peters, J. P.** Ann. Rev. Physiol. 4: 89. 1942.

rapid turnover of cerebrospinal fluid so that complete equilibrium is not attained.

It will be of some value, before proceeding farther, to examine some of the potentialities and limitations of the experimental preparation. The ligation of the renal pedicles makes available a reasonably normal animal in which the compound under study is contained in a closed system. It is to be expected that a dynamic equilibrium will be established between the concentration of the unbound substance in the fluid of reference, i e , plasma water, and its concentration in the other compartments of body fluid. The elimination of renal excretion may be expected to facilitate studies on the distribution of compounds in that it stabilizes the equilibrium concentrations that are achieved.

A change in the plasma concentration of a substance subsequent to the attainment of an initial equilibrium under these conditions is an indication that a change in the chemical structure of the compound has occurred, or, that there has been a change in the biological activity of some of the large groups of cells in the body The latter of these two factors need not complicate the interpretation of the data in such a situation since it may be excluded by properly planned experimental procedures It is to be expected in the case of the sulfonamides that the free aryl amino groups will become progressively conjugated. It is predictable that this change in chemical structure will produce a general lowering of the concentration of the substance in the plasma and in all tissues but not in the plasma/tissue distribution ratios of the parent compound In so far as a chemical modification does occur, however, it would be surprising if the resulting compound were characterized by a distribution in the body that was identical to that of the parent compound It should be possible, therefore, to use the preparation in a limited fashion in the exploration of the extent to which a change in chemical constitution occurs subsequent to the introduction of a compound into the body, to follow the rate of the change and, in certain cases, to obtain some indication of the nature of the chemical modification The use of the dog somewhat simplifies the experimental conditions because this species does not acetylate compounds of the present series to any considerable extent The evidence obtained relative to this aspect of the study will not receive extended consideration at this time The experiments selected for presentation are limited to those of two hours' duration. This interval is ample for the establishment of a stable diffusion equilibrium within the body, but is too short for the study of changes in distribution resulting from a change in the chemical structure of the compound under investigation.

It must be appreciated, in considering the data presented below, that a precise definition of the distribution of each compound has not been achieved. The magnitude of such a problem in the case of a single compound is not small. Reference to the studies of recent years on the distribution in the body of such substances as sodium, potassium and the halogens[5] will serve to fortify this view. Moreover, it must be appreciated that the data indicating that a compound enters a cell or organ yield no information on the state of the material so contained, or on the actual concentration ratio of the substance across the cell membrane, i.e., the concentration relationships in extra- and intracellular water. It is not unlikely that a considerable proportion of the specific solute within a cell reacts with cell constituents in such a manner that a portion is removed from free solution in intracellular water

The over-all renal excretion of each compound was studied in experiments such as that summarized in TABLE 2 The experiments were on

TABLE 2

THE RENAL EXCRETION OF SULFANILAMIDE BY DOG K

– :45—800 cc Water per OS Creatinine—200 mg /K. subcut. Sulfanilamide —75 mg /K subcut	:00 – Bladder emptied 3 collection periods, 15 min. each

BLOODS AT MID-PERIOD

Creatinine			Sulfanilamide				
Plasma mg./ml.	Excretion mg./min	Glomerular filtration rate ml /min	Plasma (filterable) mg /ml	Filtered mg /min	Excreted mg /min.	Reabsorbed mg./min	Excretion ratio
0.144	6 80	47.2	0.035	1.65	0.39	1.26	0.24
0.150	6.91	46.1	0.0405	1 87	0 44	1 43	0 24
0 150	7.30	48.7	0.0454	2.21	0 55	1 66	0 25

a series of normal dogs and made use of experimental techniques that permit a general description of the mechanism involved in the renal excretion of a compound. The underlying principle of the techniques is simple. Glomerular filtration rate is measured following the administration of creatinine, the rate of glomerular filtration in milliliters per minute being equal to the dividend of the milligrams of creatinine excreted per minute and the plasma concentration of creatinine in milligrams per milliliter.[6-9] Glomerular filtration rate so determined was

[6] Shannon, J. A. Am. Jour. Physiol. 112, 405. 1935.
[7] Shannon, J A Am. Jour. Physiol. 114, 362. 1936.
[8] Van Slyke, D. D., Hiller, A., & Miller, B. F. Am. Jour. Physiol. 113: 611. 1935.
[9] Richards, A. N., Bott, P. A., & Westfall, B. B. Am. Jour. Physiol. 123: 281. 1938.

calculated to be 47.2 milliliters per minute in the first period of the experiment summarized in TABLE 2. The data on the excretion of sulfanilamide, obtained simultaneously, may be manipulated as follows: The plasma concentration of sulfanilamide during this period was 0.039 mgm. per ml., 90 per cent of which is in a filterable form. At a glomerular filtration rate of 47.2 ml. per minute, it may be calculated that 1.65 mg. of sulfanilamide was filtered each minute Since the concurrent renal excretion accounts for only 0.39 of the 1.65 mg filtered each minute, one must conclude that 1.26 mg. of filtered sulfanilamide were reabsorbed in the renal tubules The calculation throws little light on the mechanism of the reabsorptive process However, the ratio of the amount filtered to the amount excreted, in this case 0.24, yields a figure that is characteristic of the excretion of the compound by the average nephron under the conditions of the experiment. This ratio, which we designate the excretion ratio, is a useful datum in the study of the renal excretion of a series of allied compounds

Another ratio, which may be calculated from the data and which does not include a correction for the factor of plasma binding, is the sulfonamide/creatinine clearance ratio The latter ratio is important in the definition of the over-all renal excretion rates of compounds, but is less useful in indicating the mechanisms by which such excretion is accomplished

The conclusion that sulfanilamide is reabsorbed to a considerable extent is contingent on the demonstration that essentially all of the plasma sulfanilamide is filterable at the glomerulus. Such a circumstance does not hold for many of the compounds that have been examined in the present study The excretion ratio, however, containing as it does a correction for the factor of plasma binding, makes available a corrected ratio that, theoretically, is a rather precise expression of the degree to which a substance is reabsorbed or actively excreted by the renal tubules It has two limitations The ratio gives limited information on the nature of the physiological mechanisms involved and it is quite sensitive to errors in the case of substances that are extensively bound on plasma protein

Precise information on the underlying mechanisms that are responsible for renal tubular reabsorption or excretion can be obtained by varying the experimental procedures. As is the case of studies on distribution of compounds, however, a quantitative definition of the renal mechanisms involved in the excretion of a single substance is a task of major proportions This viewpoint can perhaps be best appreciated by refer-

ence to studies on the excretion of inulin and creatinine,[6-9] urea,[10, 11] glucose,[12] and phenol red.[13] These were also performed in the dog and may be taken as examples of studies that define the renal excretion of compounds in terms of the discrete mechanisms involved. The data detailed below are more in the nature of a general survey of the excretion of this class of compounds. Extraneous factors were minimized by adhering to rigidly standardized experimental conditions. These include a state of moderate diuresis, and a plasma concentration for the substance under examination of the order of magnitude of 5.0 mgm. per cent. All observations were made within 90 minutes after the administration of the compound.

It may be accepted in the following that an excretion ratio of 1.0 is indicative of a situation wherein the renal tubules do not participate in the excretion of a compound to a measurable extent. As the ratio falls progressively below 1.0, it is indicative of a progressive increase in the reabsorption of a compound in the renal tubules. A ratio above 1.0 is indicative of tubular excretion, the extent of which is reflected in the value of the excretion ratio. Tubular excretion, in the case of this series of compounds, is presumably by an active process which specifically extrudes the material into the lumen of the tubule.[14] The overall rate of excretion of a compound in the latter case is equal to the sum of its rate of filtration at the glomerulus and its rate of secretion by the tubules.

The extent to which each of the compounds is bound on the non-diffusible constituents of plasma was determined by dialysis of plasma samples using cellophane or parlodion membranes at 37° C., p. CO_2 = 38 mm Hg. The values shown in TABLES 2, 5, 10, 11, 12 and 13 are those obtained with normal dog plasma at albumin concentration of 4.0 g. per cent. A more limited series of compounds (twelve) was examined, using cat plasma. Assuming the latter to be representative of the series as a whole, it may be concluded that the figures on the plasma binding in the dog are a fair approximation of the situation obtaining in the normal cat. A more precise definition of plasma binding in the cat was not deemed essential for the purposes of the present study.

It may be accepted that within the limitations that we have noted, the above techniques permit a description of the distribution of substances in the various compartments of body water and, together with observations on plasma binding, the extent to which renal tubular

10 Shannon, J. A. Am. Jour. Physiol. 117: 206. 1936.
11 Shannon, J. A. Am. Jour. Physiol. 122: 782. 1938.
12 Shannon, J. A., & Fisher, S. Am. Jour. Physiol. 122: 765. 1938.
13 Shannon, J. A. Am. Jour. Physiol. 113: 602. 1935.
14 Shannon, J. A. Physiol. Rev 19: 63. 1939.

processes participate in the determination of the over-all excretion of each compound. The data made available are wholly satisfactory for our present purpose, which is to survey the distribution and excretion of a series of compounds preliminary to more definitive studies on selected members within the series.

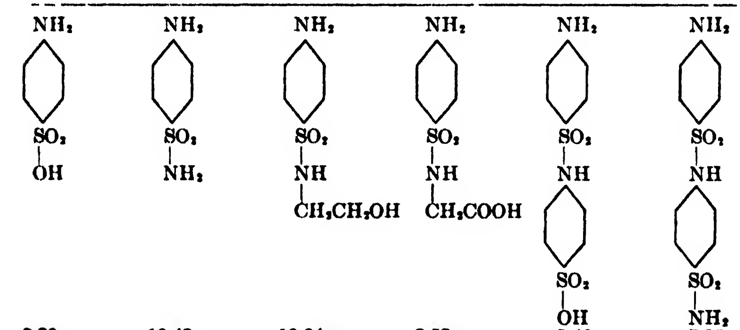
EXPERIMENTAL RESULTS

Studies on the distribution and excretion of six selected compounds have been summarized in TABLES 3, 4 and 5. These demonstrate some of the salient features of the data as a whole and so may serve for orientation purposes. TABLE 3 gives the structural relationships of the

TABLE 3

STRUCTURAL RELATIONSHIPS OF THE SIX SULFONAMIDES SELECTED FOR PRELIMINARY EXAMINATION

Sulfanilic acid	Sulfanilamide	Sulfanilylethanolamide	Sulfanilylglycine	Sulfanilylsulfanilic acid	Sulfanilylsulfanilamide
3.20	10 43	10.94	3.52	3.40	7 85



series and the pKa of each compound. The simplest view one may take of the series is that the five compounds to the right are simple derivatives of sulfanilic acid.

TABLE 4 summarizes the distribution data on each of the six compounds in a few selected tissues of the cat two hours subsequent to its intravenous administration. The values are, in all cases, the means of those obtained in several experiments. The distribution of sulfanilic acid appears to be characteristic of a substance that is largely restricted to an extracellular position, since the volume of distribution and the

TABLE 4

SUMMARY OF OBSERVATION ON THE DISTRIBUTION IN THE CAT OF SIX SELECTED SULFONAMIDES

Compound	Ratio: concentration in tissue / concentration in plasma/0.93			Volume of distribution	pKa
	Muscle	CSF	Brain	% Body Weight	
Sulfanilic acid	0.11	0.03	0.03	28.7	3.20
Sulfanilamide	1.07	0.68	0.73	98.2	10.43
Sulfanilylethanolamide	0.70	0.30	0.21	67.2	10.94
Sulfanilylglycine	0.16	0.03	0.04	29.9	3.52
Sulfanilylsulfanilic acid	0.14	0.02	0.04	35.7	3.40
Sulfanilylsulfanilamide	1.23	0.07	0.15	165.0	7.85

distribution ratios are of the same order of magnitude as those obtaining for sodium or chloride. In addition, it appears to be unable to pass the blood-brain barrier with any degree of freedom. Note may be taken here of the fact that this appears to be as true for the capillary plexes in brain substance as for those that are responsible for the formation of cerebrospinal fluid, i.e., if these be different anatomical structures.[15] Such a distribution is completely changed both with respect to tissues in general and the blood-brain barrier, when one blocks the dissociation of the sulfonic acid by an amino group as in the case of sulfanilamide. Presumably this occurs because of a change in the physicochemical characteristics of the nuclear substituent. The new compound not only freely enters the cell and readily crosses the blood-brain barrier, but is localized to some extent within the cells of the body The distribution ratio for muscle[16] would otherwise be in the order of magnitude of 0.75.

Further elaboration of the substituent by the addition of an ethanol does not significantly change the distribution across cell membranes in general, although it does appear to impair the ability of the resulting compound to enter cerebrospinal fluid. The oxidation of the alcohol to an acid, however, returns the distribution to the type that, as in the case of sulfanilic acid, is largely extracellular. The substitution of an amide hydrogen by benzene sulfonic acid again results in a compound that is largely localized in an extracellular position, whereas blocking the terminal sulfonic acid group of the latter compound by an amino group permits the resulting compound to pass freely into intracellular water. Were a correction made for the factor of plasma binding in the

[15] Wallace, G. B., & Brodie, B. B. Jour. Pharm. and Exp. Therap. 70: 418. 1940.
[16] Skelton, H. Arch. Int. Med. 40: 140. 1927.

calculation of the distribution ratios, it would be apparent that sulfanilylsulfanilamide is localized within the cells to a much greater extent than is true in the case of sulfanilamide itself. This follows from the fact that about half of the plasma sulfanilylsulfanilamide is bound on plasma protein. The true equilibrium concentration is of course the concentration in extracellular water, and not the concentration in plasma as a whole.

Sulfanilylsulfanilamide is of interest because of still another reason. Accepting that only 50 per cent of the plasma concentration is available to establish a diffusion gradient, it is apparent that the compound is incapable of freely penetrating the blood-brain barrier. It is quite possible that the diffusion of the complex across the barrier is simply slow and equivalent concentrations would eventually be reached were the rate of turnover of cerebrospinal fluid curtailed. It may be concluded, however, that the substitution has resulted in a compound that does not penetrate the barrier as readily as sulfanilamide, that this property of the complex is reflected in a low concentration of the substance in brain substance, and that this property of the complex is not related specifically to the acidic properties of the nuclear substituent

Another substance appearing in TABLE 4 is of interest from a different aspect. This is the sulfanilylethanolamide which appears to be completely changed in the body and which may serve as an example of how such a circumstance may be examined. The distribution at two hours, as noted in TABLE 4, is quite similar to that of sulfanilamide. With the elapse of time, however, the material progressively acquires a distribution that is largely extracellular. The change in the distribution of the compound is already apparent at 6 hours and almost complete at 24 The distribution at 48 hours is identical to that of sulfanilylglycine. It is not possible to conclude from such data that the chemical transformation is in fact an oxidation of the alcohol to the acid, although this does seem quite probable. Equally extensive changes have been observed in the case of several other compounds, but since this aspect of the problem is not germane to the present discussion, they will not be cited in the detailing of the data below

Excretion data on the selected series are reviewed in TABLE 5. It will be recalled that a progressive decrease in the excretion ratio below 1.0 indicates a progressive increase in the reabsorption of the compound during its passage down the renal tubules, whereas a progressive increase in the excretion ratio above 1 0 indicates tubular excretion which proceeds in proportion to the elevation of the ratio. The data on sulfanilic acid indicate that it is excreted largely by a process of glom-

TABLE 5

SUMMARY OF OBSERVATIONS ON THE RENAL EXCRETION OF SIX SELECTED SULFONAMIDES BY THE DOG

Compound	pKa	Sulfonamide / Creatinine clearance ratio	Per cent filterable in plasma	Excretion ratio
Sulfanilic acid	3.20	0.89	98	0.91
Sulfanilamide	10.43	0.27	90	0.30
Sulfanilylethanolamide	10.94	0.63	89	0.71
Sulfanilylglycine	3.52	1.35	92	1.47
Sulfanilylsulfanilic acid	3.40	2.23	34	6.29
Sulfanilylsulfanilamide	7.85	1.34	56	2.39

erular filtration, whereas sulfanilamide is reabsorbed to a considerable extent. The addition of an ethanol to the sulfanilamide sharply curtails this reabsorption, although some reabsorption still is evident But the oxidation of the alcohol to the acid produces a molecular complex, which, in addition to being filtered at the glomerulus, is actively secreted by the renal tubules. The addition of benzene sulfonic acid to the sulfanilamide molecule further enhances such tubular excretion and this is retained as a property of the compound that results when the dissociation of the sulfonic acid in the latter compound is blocked by the addition of an amino group.

It will be profitable to pause at this time, in order to consider some of the generalities apparent from the data so far presented. It has been demonstrated that a change in chemical structure within this small series of compounds is accompanied by drastic consequences to those factors that together determine the physiological disposition of a compound. Were one to assume some common chemotherapeutic effect on the part of all six compounds, it is doubtful that sulfanilic acid or compounds comparable to it from the standpoint of distribution would exert such an action were the site of action within the central nervous system, unless there were some change in the characteristics of the blood-brain barrier This follows from the fact that compounds of the latter type are excluded from this compartment of the body by the normal blood-brain barrier. Similarly, one could reason that if it were desirable to obtain a temporary high concentration of one of these in the extracellular fluid compartment in order to accomplish a given end, then the compound of choice would be either sulfanilic acid, sulfanilylglycine, or sulfanilylsulfanilate. This follows from the limitation imposed on the distribution of the compounds together with their relatively high excretion rates, the latter largely determining the dura-

tion of the desired effect. For longer action, one would select the compound with the lower excretion rate. If one were not concerned with a limitation on the distribution of a compound and wished an effect with considerable duration, then one would select a compound such as sulfanilamide which has a large volume of distribution together with a low excretion rate.

In other words, there is some justification for the belief that it may ultimately be possible to limit both the theatre of operation of an active substance within a series as well as the duration of the effect produced through the proper modification of chemical structure. The data on these six compounds do not lead one to suppose that this will be a simple procedure. Wholly apart from the effect of a chemical change on fundamental activity of the resulting compound, at least two factors will be concerned with the distribution and excretion. The strength of acidic groups is one of these, whereas the other apparently relates to the molecular structure of the compound as a whole The latter factor is difficult to define in terms of measurable entities at the moment, but is none the less important Its operation is particularly apparent from a consideration of the excretion data on sulfanilamide as compared to sulfanilylsulfanilamide. The operation of this factor is even more apparent when the data of the entire series are examined

The distribution data on a series of acyclic derivatives of sulfanilamide and of a series of closely allied compounds are summarized in TABLES 6 and 7. There are several points of interest in the series of

TABLE 6

SUMMARY OF OBSERVATIONS ON THE DISTRIBUTION IN THE CAT OF A SERIES OF N^1-ACYCLIC DERIVATIVES OF SULFANILAMIDE

−R	Ratio: concentration in tissue / concentration in plasma/0.93			Volume of distribution	pKa
	Muscle	CSF	Brain	% Body weight	
$-H$	1 07	0 68	0 73	98.2	10.43
$-CH_3$	1.02	0 77	0 85	91 7	10 77
$-CH_2CH_3$	1.11	0 89	0 92	102.0	10.70
$-OH$	1 07	0.58	0.60	99.0	—
$-NH_2$	—	0.43	—	72.5	10.70
$-CH_2CH_2OH$	0 70	0.30	0 21	67.2	10.94
$-C{=}NH\cdot NH_2$	0.51	0.09	0.08	78.6	Very Weak
$-CH_2COOH$	0.16	0.03	0.04	29.9	3.52
$-COCH_3$	0.37	0.05	0.06	55.5	5.38

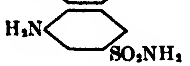
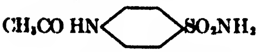
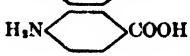
acyclic derivatives. The addition of a non-polar group to the parent sulfanilic acid has, in general, produced a substance that penetrates cell membranes of all tissues and that, in addition, is localized within the contents of most cells. The N^1-methyl and N^1-ethyl derivatives of sulfanilamide appear to enter cerebrospinal fluid more freely than sulfanilamide, but since sulfanilamide itself passes the blood-brain barrier so readily, the difference is not striking. But, the hydroxy, amino, and ethanol derivatives unquestionably show an impairment in this ability which is quite striking in the case of the sulfanilylguanidine. The data on the N^1-hydroxysulfanilamide are not very satisfactory since a fairly rapid reversion of the compound to sulfanilamide is to be expected and, consequently, an eventual distribution in the body which is characteristic of the latter substances.

The N^1-acetyl derivative of sulfanilamide and sulfanilylglycine have quite a different distribution, being somewhat similar to that of sulfanilic acid, presumably because they are in fact organic acids with a strength of the same order of magnitude as that compound. This appears to be reflected in the degree to which all three are excluded from cerebrospinal fluid and to some extent from cellular water in general. The correlation between over-all distribution and acidic strength is crude, however, as may be seen from examination of the last two columns. Inspection of specific tissues, i e., muscle *vs* cerebrospinal fluid, fortifies the belief expressed previously that distribution is determined in part by unspecified but general characteristics of a molecule as well as by the charge carried on a specific group or its tendency to dissociate a hydrogen ion.

Data on some compounds closely allied to sulfanilamide are given in TABLE 7. The data on para-aminobenzoic acid are particularly worthy of note. One would expect from inspection of the molecule that its distribution would not be very different from the distribution of sulfanilylglycine or that it would lie midway, in its distribution, between the latter compound and perhaps that of the N^1-acetyl derivative of sulfanilamide. The correlation is not too bad in the case of the volume of distribution and the distribution ratio of muscle, but it diffuses into cerebrospinal fluid quite as easily as sulfanilamide itself. It is important to note, taking para-aminobenzoic acid as a specific example, that the distribution ratio is not necessarily an expression of the concentration ratio across the cell membrane. Assuming that this substance is of importance to the economy of the living cell, it would be surprising were it not localized within the cell in such a manner as to

TABLE 7

SUMMARY OF OBSERVATIONS ON THE DISTRIBUTION IN THE CAT OF A SERIES OF SIMPLE COMPOUNDS CLOSELY ALLIED TO SULFANILAMIDE

Compound	Ratio: $\frac{\text{concentration in tissue}}{\text{concentration in plasma/0.93}}$			Volume of distribution	pKa
	Muscle	CSF	Brain	% Body weight	
	1.07	0.68	0.73	98.2	10.43
	—	0.76	—	103.0	10.12
	1.00	0.71	0.84	92.3	10.12
	0.97	0.27	0.15	84.2	—
	0 11	0 03	0.03	28 7	3.20
	0.40	0.64	0.32	51.4	4.68

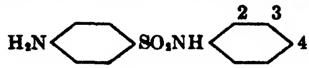
make some of the cellular para-aminobenzoic acid unavailable for conditioning its rate of diffusion.

It is of some interest to note that the N[4]-acetyl derivative of sulfanilamide has less restriction imposed on its distribution in the body than when the acetyl group is in the N[1] position. The simplest view of this is that the former is a weaker acid The divergence in distribution in the body as a whole, however, and the distribution across the blood-brain barrier suggests that other factors operate in conditioning the distribution of the compound.

A summary of the distribution of ten isocyclic derivatives of sulfanilamide is given in TABLE 8. Again, the presence or absence of a reasonably strong acidic group would appear to be important in determining the ability of a substance to penetrate cell membranes in general, as evidenced by the volume of distribution of a compound when viewed in relation to its pKa. The situation with respect to these compounds, however, is not as simple as would appear from these data. It should be noted that a modification of the molecular structure to include a substituted or nonsubstituted benzene ring throws into prominence the ability of the nondiffusible constituents of plasma to form reversible combinations with the resulting compounds (see TABLES 10

TABLE 8

SUMMARY OF OBSERVATIONS ON THE DISTRIBUTION IN THE CAT OF A SERIES OF N^1-ISOCYCLIC DERIVATIVES OF SULFANILAMIDE



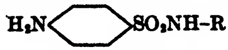
Substituent	Ratio: concentration in tissue / concentration in plasma/0.93			Volume of distribution	pKa
	Muscle	CSF	Brain	% Body weight	
—	0.80	0.18	1.06	125.0	9.60
4-NH_2	1.15	0.33	0.51	124.0	10.22
3-SO_2NH_2	0.94	0.14	0.19	146.0	8.23
4-SO_2NH_2	1.23	0.07	0.15	165.0	7.85
2-COOH	—	0.01	—	31.6	3.85
3-COOH	0.07	0.01	0.03	23.0	4.10
4-COOH	0 15	0.01	0.04	36.2	4.05
2-SO_2OH	0.10	0.01	0.06	23.9	3.40
3-SO_2OH	0.08	0.01	0.04	18.8	3.35
4-SO_2OH	0.14	0.02	0.04	35.7	3.40

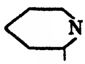
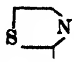
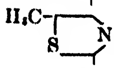
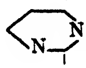
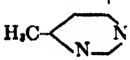
to 13. It would be apparent, were a correction made for this factor, that the tissue concentration of several compounds in the series is far in excess of the amount that would obtain were it simply related to the concentration of the compound in extracellular fluid and the concentration of water in the tissue. The data on the N^1-phenyl, the N^1-4 aminophenyl and the N^1-sulfonamidophenyl derivatives indicate a very extensive localization of each of these substances within the cells. The blood-brain barrier again appears to have specific characteristics, however, that set cerebrospinal fluid and interstitial fluid of the brain apart from the interstitial fluid of the other organs of the body. The best example of this in the present series is sulfanilylsulfanilamide, as has been noted above.

The data on the distribution of a series of N^1-heterocyclic derivatives of sulfanilamide are summarized in TABLE 9. The degree to which each is bound on plasma protein should also be kept in mind in the examination of these data. The order of magnitude for each of the five being 30 per cent for sulfapyridine, 60 per cent for sulfathiazole, 80 per cent for the 4-methyl, sulfathiazole (sulfamethylthiazole), 20 per cent for sulfadiazine, and 40 per cent for the 4-methyl sulfadiazine (sulfamerazine). It is possible to conclude with this information that there is actual localisation within the cell in the case of sulfapyridine, sulfa-

TABLE 9

SUMMARY OF OBSERVATIONS ON THE DISTRIBUTION IN THE CAT OF A SERIES OF N^1-HETEROCYCLIC DERIVATIVES OF SULFANILAMIDE



−R	Ratio: $\frac{\text{concentration in tissue}}{\text{concentration in plasma/0.93}}$			Volume of distribution	pKa
	Muscle	CSF	Brain	% Body weight	
	0.91	0.62	0.80	82.5	8.44
	0.54	0.10	0.14	58.5	7.12
	0.46	0.13	0.16	51.4	7.80
	0.45	0.31	0.21	46.0	6.48
	0.39	0.38	0.35	45.8	7.06

thiazole and sulfamethylthiazole, whereas sulfadiazine and sulfamerazine are distributed more as if they freely penetrate most cell membranes, and are not specifically localized within the cells. The spread in the dissociation constants in the series is not great, and over the limited range there does not appear to be any important correlation between them and the distribution of the compounds.

Data bearing on the renal excretion of the compounds in the series have been summarized in TABLES 10 to 13. The form of presentation of each group is similar to that in TABLE 10. Each value is the mean of several experiments performed on at least two dogs. Information is given on the sulphonamide/creatinine clearance ratio that describes the over-all rate of renal excretion, the fraction of the sulfonamide which is filterable from the plasma at a normal concentration of plasma albumin, the pKa of the compound, and its excretion ratio.

The addition of a relatively non-polar group to the sulfonamide complex may enhance or depress the excretion of a compound to a variable extent (TABLE 10). Such changes in excretion are largely the result of a change in the extent to which the compound is reabsorbed, since this class of compounds does not appear to be bound extensively on plasma protein. When a strongly polar group with acidic proper-

TABLE 10

SUMMARY OF OBSERVATIONS ON THE EXCRETION BY THE DOG OF A SERIES OF N¹-ACYCLIC DERIVATIVES OF SULFANILAMIDE

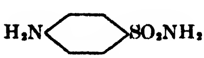
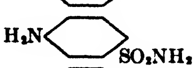
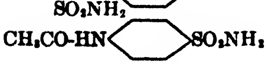
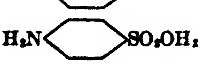
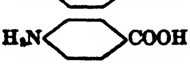
−R	pKa	Sulfonamide / Creatinine clearance ratio	Per cent filterable in plasma	Excretion ratio
$-H$	10.43	0.27	90	0.30
$-CH_3$	10.77	0.12	80	0.15
$-CH_2CH_3$	10.70	0.11	79	0.14
$-OH$	—	0.35	88	0.40
$-NH_2$	10.70	0.43	83	0.52
$-CH_2CH_2OH$	10.94	0.63	89	0.71
$-C{=}NH \cdot NH_2$	Very weak	0.76	94	0.81
$-CH_2COOH$	3.52	1.35	92	1.47
$-COCH_3$	5.38	1.19	87	1.37

ties is added, however, it may convert the compound into a complex that is actively secreted by the renal tubules as is the case with sulfanilylglycine and the N¹-acetyl derivative of sulfanilamide.

It does not appear, however, that any generalization may be derived from these data since the results are not wholly consistent with the information obtained on a series of simple allied compounds It is apparent from the latter data (TABLE 11) that sulfanilic acid is reabsorbed

TABLE 11

SUMMARY OF OBSERVATIONS ON THE EXCRETION BY THE DOG OF A SERIES OF SIMPLE COMPOUNDS CLOSELY ALLIED TO SULFANILAMIDE

Compound	pKa	Sulfonamide / Creatinine clearance ratio	Per cent filterable in plasma	Excretion ratio
	10.43	0.27	90	0 30
	10.12	0.17	73	0.23
	10.12	0.25	89	0.28
	—	0.67	80	0.84
	3.20	0.89	98	0.91
	4.68	0.18	93	0.19

to a slight extent and a similarly strong organic acid, para-aminobenzoic acid, is reabsorbed to a considerable extent. The excretion data on para-aminobenzoic acid are particularly interesting in that it has a dissociation constant midway between that of sulfanilylglycine and the N^1-acetyl derivative of sulfanilamide. But para-aminobenzoic acid is reabsorbed to an extent that suggests the intervention of an active tubular process, whereas sulfanilylglycine and the N^1-acetyl derivative of sulfanilamide are both actively excreted by the renal tubules.

Other data in TABLE 11 are of interest in that they show that metanilamide and orthanilamide are handled by the renal nephron in a manner that is quantitatively similar to that of sulfanilamide itself. This would indicate that the position of the substituent groups is of little importance in determining the excretion of the members of this family of three as it was of little consequence in determining their distribution in the body. It is of some interest to note that whereas the N^1-acetyl derivative is actively excreted, the N^4-acetyl derivative is excluded from extensive participation in renal tubular processes within the nephron. It is also worthy of comment that even as the glycine derivative of sulfanilic acid is actively excreted by the renal tubules, so is the glycine derivative of para-aminobenzoic acid, i.e., para-aminohippuric acid [17]

The data on the renal excretion of some of the isocyclic derivatives of sulfanilamide are summarized in TABLE 12. It may be noted that such compounds are, in general, bound extensively on the nondiffusible constituents of plasma; that the N^1-phenyl is itself actively secreted, whereas the $N^1$4-aminophenyl is reabsorbed to a considerable extent; that the sulfonic acid derivatives of N^1-phenyl sulfanilamide are actively secreted as are the carboxy derivatives; but that although sulfonamido derivatives are actively secreted when the nuclear substituent is in the 3 or 4 position, this is not true when the substituent is in the 2 position. There is also a suggestion that the position of the substituent group has some effect upon the rate of active transport in the group of sulfonic acid derivatives The excretion ratio in this series increases progressively from 6 to 14 as the sulfonic acid is rotated from the 4 to the 3 to the 2 position. The phenomenon is not apparent, however, in the case of the carboxy derivatives It is not surprising to find that the position of a substituent may have an effect upon the extent to which a compound is bound on plasma protein or upon the rate of a biological process that involves a series of chemical combinations.[14] It is somewhat surprising that the position can determine whether a

[17] Finkelstein, N., Aliminosa, L. M., & Smith, H. W. Am. Jour. Physiol. 133: 276. 1941.

TABLE 12

SUMMARY OF OBSERVATIONS ON THE EXCRETION BY THE DOG OF A SERIES OF N^1-ISOCYCLIC DERIVATIVES OF SULFANILAMIDE

Substituent	pKa	Sulfonamide / Creatinine clearance ratio	Per cent filterable in plasma	Excretion ratio
—	9.60	1.00	28	3.57
4-NH_2	10.22	0.36	65	0.55
2-SO_2NH_2	—	0.08	35	0.22
3-SO_2NH_2	8.23	0.73	38	1.92
4-SO_2NH_2	7.58	1.34	56	2.39
2-COOH	3.85	0.77	12	6.42
3-COOH	4.10	1.84	15	12.26
4-COOH	4.05	1.67	34	4.91
2-SO_2OH	3.40	2.22	16	13.90
3-SO_2OH	3.35	2.30	24	9.58
4-SO_2OH	3.40	2.23	34	6.29

substance is actively secreted or, as appears to be likely in the case for the sulfanilylorthanilamide, actively reabsorbed.

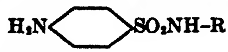
The excretion data from this series of compounds are important for another purpose. They oppose the acceptance of the concept that the participation or nonparticipation of a compound in processes of active tubular transfer is determined largely by the presence of so-called organophilic and hydrophilic groups that, arranged across the molecule, serve to orient it with respect to the cell membrane.[18] Such an orientation has been supposed to be essential for the entry of a substance into the renal tubule cell prior to its active transport across the cell. It is unquestionably true that the participation or nonparticipation of a substance in an active tubular mechanism will depend in part upon the physicochemical characteristics of the molecule as a whole, but the manner in which this factor operates is not sufficiently clear at the moment to permit definitive conclusions on the mechanism involved.

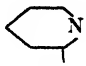
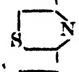
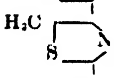
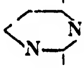
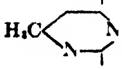
The excretion data on some of the important heterocyclic derivatives of sulfanilamide are summarized in TABLE 13. The series is small, but the data contain several points of considerable interest. They indicate that sulfapyridine, sulfadiazine and sulfamerazine are reabsorbed

[18] Hoeber, R. Cold Spring Harbor Symp. Quant. Biol. 8: 40. 1940.

TABLE 13

SUMMARY OF OBSERVATIONS ON THE EXCRETION BY THE DOG OF A SERIES OF N^1-HETEROCYCLIC DERIVATIVES OF SULFANILAMIDE



−R	pKa	Sulfonamide / Creatinine clearance ratio	Per cent filterable in plasma	Excretion ratio
	8.44	0.38	69	0.55
	7.12	0.40	40	1.00
	7.80	0.25	23	1.09
	6.48	0.27	83	0.33
	7.06	0.13	61	0.21

to a considerable extent in the renal tubules, whereas sulfathiazole and sulfamethylthiazole appear to be excreted by a process that is almost exclusively glomerular filtration. These conclusions follow from the values of the excretion ratios given in the last column of the table which in the case of the thiazoles approximate a value of 1.0.

Reabsorption is considerable in the case of sulfamerazine, i.e., 4-methyl sulfadiazine and, as in the case of para-aminobenzoic acid, is of an order of magnitude that suggests that reabsorption proceeds as an active tubular process. When it is appreciated that a diffusible substance such as urea is only reabsorbed to the extent of some 40 per cent at ordinary urine flows, it is surprising that organic complexes of the size and shape of sulfamerazine and para-aminobenzoic acid are reabsorbed to the extent indicated without the intervention of an active process.

DISCUSSION

The study as a whole should not be taken as a definitive description of the relationship between the chemical structure of these compounds and their physiological disposition. It is at best a survey indicating the worthwhileness of this approach to the problem.

The results on the distribution of the compounds in the body are fairly consistent in that the ability of a compound to pass a cell membrane appears to be governed to a large degree by the extent of its dissociation at blood pH. Those compounds with a high pKa are permitted to enter the tissue cells quite freely, whereas those with a low pKa have in general a more restricted distribution. In this respect, therefore, sulfonamides, with some exceptions, follow the accepted concepts of permeability of cell membranes to organic electrolytes.[19]

The observations on cerebrospinal fluid offer additional proof that the barrier between the blood and cerebrospinal fluid is a highly selective one, and that the mechanism of formation of this fluid is not one of simple filtration. It has been shown previously in studies on the distribution of thiocyanate, bromide and iodide, that cerebrospinal fluid is formed, not only at the choroid plexus, but also by the passage of fluid and solute from blood into the extracellular fluid of the brain and then into the subarachnoid space.[15] The data are wholly consistent with the view that brain tissue must be considered to be in equilibrium with a fluid having the characteristics of cerebrospinal fluid rather than the characteristics of plasma water. The behavior of the normal blood-brain barrier is quite different from that of the ordinary cell membranes in that the molecular configuration of a substance, apart from its acidic properties, seems to be important in determining the ability of a substance to gain entrance into cerebrospinal fluid and the interstitial fluid of brain substance. The brain capillaries and their surrounding pia do not offer a satisfactory anatomical explanation of such selectivity.

The relation between the chemical structure of organic compounds and their ability to participate in the active excretory mechanisms of the renal tubule has been the subject of previous investigations by others. A large number of dyes as well as several sulfanilamide derivatives have been examined, using the perfused frog kidney as an indicator. It was concluded from these studies that the polarity of the molecule largely determined whether a compound participated or did not participate in a mechanism of active tubular excretion Those compounds with polar–non-polar or so-called organophilic-hydrophilic configuration appeared to be secreted, while non-polar molecules did not.[18] Many of the actively secreted compounds in the present series possess such a polar–non-polar configuration, but there are several striking exceptions. For example, sulfanilylsulfanilamide is actively secreted, although it contains two weakly polar groups, an amino and a sulfanilamido group. Sulfanilic acid and para-aminobenzoic acid

[19] Jacobs, M. H. Cold Spring Harbor Symp. Quant. Biol. 8: 40 1940.

are strongly polar molecules and are not secreted, whereas comparable compounds such as the sulfanilylglycine, the N^1-acetyl sulfanilamide and para-aminohippuric acid are. Molecular configuration is undoubtedly important in determining the participation or nonparticipation of a compound in a process of active tubular excretion, but the mechanisms that are involved require further clarification.

THE TOXIC EFFECTS OF SULFONAMIDES *

By H. B. van Dyke

From the Division of Pharmacology,
The Squibb Institute for Medical Research,
New Brunswick, N. J.

Discussion of the toxic actions of sulfonamides requires reference to a few general principles of pharmacology, so that the reader unfamiliar with this field may be better oriented with respect to terminology and the implications of this terminology. One fundamental law of the action of a drug or poison is that this action leads to no new function on the part of a tissue or organ. The group of signs or symptoms, which might appear bizarre, unusual, or unique, is always found on analysis to represent an abnormal combination of a group of simultaneous quantitative changes sometimes modified by known or unknown factors, such as heredity or disease. The precision of measurement of toxicity naturally varies greatly. The more completely the observer can exercise control over the conditions of examination, the greater the precision will be. Conditions of control are most favorable in the experimental laboratory and it is natural, consequently, that the most accurate measurements of a given toxicological feature or group of features of a drug or poison can be made in the laboratory. Laboratory animals no less than men vary in their response to drugs and poisons Therefore, it is necessary that at least a simple application of the laws of probability be employed in evaluating results so obtained, provided that the number of individuals investigated justifies such mathematical treatment The pioneer work of Trevan[1] called attention to the necessity of such treatment of experimental data. His work has been extended and more elaborate methods of the treatment of data have been evolved by biologists and statisticians, such as Gaddum, Fisher and Bliss By the use of such statistical techniques it is possible to relate the probability of error with the magnitude of error under a particular set of conditions. It is important to recognize that information so obtained has no bearing on the accuracy of the determination. Since the dose of a drug or poison related to an event, such as death, is commonly

* In this review, the author has attempted to describe the important toxic features of sulfonamide therapy without undertaking an exhaustive citation of the large number of published observations. The available literature to May 1943, was reviewed

[1] Trevan, J. W. Proc. Roy Soc. Series B. 101: 483. 1927.

related in a linear fashion to that event when about 50 per cent of the animals are affected, best agreement among experiments in a given laboratory or among different laboratories is obtained if the dose responsible for a given effect is expressed as that causing the effect in 50 per cent of a group of animals. Moreover, when the proportion of animals affected is about 50 per cent, a given increment of dose usually produces a greater change than at higher or lower dose-levels.

It is of interest to point out that sulfonamides as a group represent drugs that are administered in a very high dose and are employed in amounts enormously greater than those of very potent drugs. For example, the initial oral dose of digitoxin with a molecular weight of 764 is one mg, whereas the initial oral dose of sulfanilamide with molecular weight of 172 is about four grams. Thus, by actual weight, 4000 times as much sulfanilamide as digitoxin is administered. The discrepancy is still more striking if one calculates dosage in terms of molecules of drugs. Under these circumstances the dose of sulfanilamide is nearly 18,000 times as great as that of digitoxin

The immediate toxic effect ("acute toxicity") of sulfonamides is of limited value in predicting toxic potentialities It is usually essential that such a determination be made in comparison with known compounds under rigidly standardized conditions. The degree of acute toxicity will depend upon the route of administration and may be vastly different if one compares oral doses with doses injected subcutaneously, intraperitoneally or intravenously. Important factors in such a comparison are often the solubility and absorbability of the sulfonamide used. For example, Powell and Chen[2] were not able to administer enough sulfapyridine by stomach tube to kill 50 per cent of a group of mice. Apparently the drug was so insoluble and so little absorbed that the capacity of the stomach limited dosage to the equivalent of 34 grams per kilo body weight However, under the best of conditions the information obtained is only of preliminary interest. For example, the chief symptoms exhibited by animals receiving an acute lethal dose of a sulfonamide are referable to the central nervous system The investigator may observe ataxia, spastic paralysis, tonic and clonic convulsions and other signs and symptoms often culminating in coma Such signs are almost never observed in patients receiving sulfonamides.

An example of an early determination of the acute toxicity of sulfanilamide is given in TABLE 1, reproduced from the article of Marshall,

[2] Powell, H. M., & Chen, K. K. Jour. Pharmacol. 67: 79. 1939.

TABLE 1

THE TOXICITY OF SULFANILAMIDE AND ACETYLSULFANILAMIDE IN MICE
(Marshall, Cutting and Emerson[3])

	Sulfanilamide		Acetylsulfanilamide	
Dose Gm./kg.	Number tested	Per cent mortality	Number tested	Per cent mortality
1	12	8.3	12	0.0
2	28	7.1	20	45.0
3	24	25.0	23	57.0
4	20	65.0	20	70.0
5	..	....	12	75.0
6	24	88.0	24	92.0

Cutting and Emerson. FIGURE 1 illustrates the acute toxicity of a number of sulfonamides as determined by Walker and van Dyke.[4] It is of interest that the data of Marshall, Cutting and Emerson[3] indicate that acetylsulfanilamide is more toxic than sulfanilamide itself. From FIGURE 1 it can be concluded that the toxicity of sulfathiazole derivatives varies in a manner that could not be predicted and that actual experiment alone demonstrated that sulfaethylthiazole is the most toxic of the sulfonamides examined in this group. These two examples furnish information concerning the possible utility of determination of the acute lethal dose of sulfonamides.

Marshall and his co-workers were pioneers in the investigation of

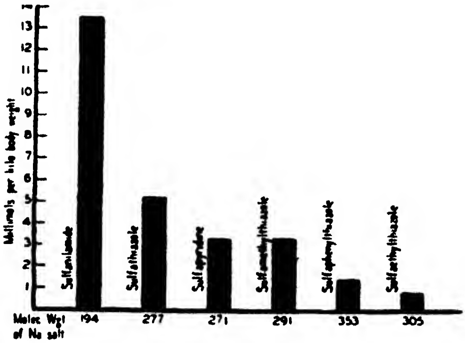


FIGURE 1. Median lethal dose (L. D. 50) of various sulfonamides after one subcutaneous injection of sodium sulfonamide into mice. To estimate the L. D. 50, groups of mice not smaller than 20 each were used (from Walker and van Dyke[4])

[3] Marshall, E. K., Jr., Cutting, W. C., & Emerson, K. Jour. Am. Med. Assoc. 110: 252. 1938.
[4] Walker, H. A., & van Dyke, H. B. Jour. Pharmacol. 71: 138. 1941.

the relationship between the blood level of sulfonamides and toxic or therapeutic effects. In FIGURE 2, from the publication of Marshall, Cutting and Emerson,[3] the blood levels in relation to dose and compound are illustrated with regard to sulfanilamide and acetylsulfanilamide. In general, acetyl derivatives of sulfonamides are less soluble, less readily absorbed, and more toxic than the corresponding free acids. However, these statements do not necessarily apply to some of the methyl substitution derivatives of sulfadiazine.

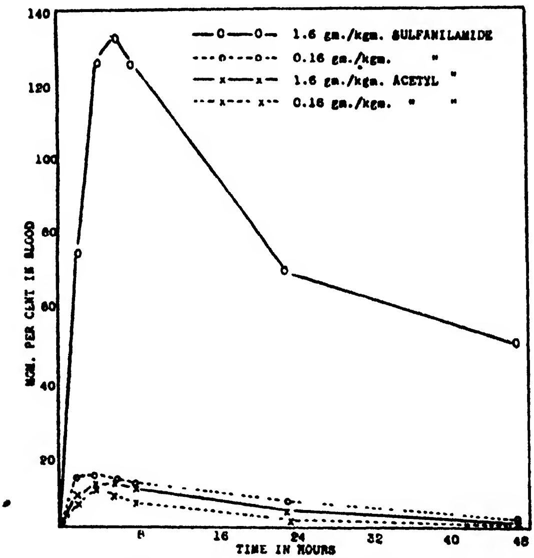


FIGURE 2 Difference in absorption of sulfanilamide and acetylsulfanilamide in small and large dosage in a dog given at different times a small and a large dose of these drugs. Dosage and concentration of acetylsulfanilamide are expressed in terms of sulfanilamide (from Marshall, Cutting and Emerson[3])

Acetylation takes place principally in the liver, which presumably employs acetate derived from pyruvate or acetaldehyde, or both. The feeding of acetate or pyruvate or acetoacetate increases the degree of acetylation of an aromatic amino compound like para-aminobenzoic acid (Hensel[5]). Other experiments have been performed by Harrow, Power and Sherwin.[6] Recent experiments on the mechanism of acetylation employing rabbit liver tissue and sulfanilamide have been re-

[5] Hensel, M. Zeit. physiol. Chem. 93: 401 1915.
[6] Harrow, B., Power, F. W., & Sherwin, C. P. Proc. Soc. Exp. Biol Med. 24: 422 1926.

ported by Klein and Harris.[7] They concluded that intact liver cells alone can conjugate acetate and sulfonamide. This reaction is irreversible and its rate is only slightly reduced under anaerobic conditions. The authors also discussed the formation of pyruvate from lactate as well as the conversion of pyruvate or acetaldehyde into acetate

Marshall, Cutting and Emerson[7] showed that the following relationship between blood level and symptoms of toxicity could be demonstrated in dogs receiving sulfanilamide:

30 mg. per cent in blood	no symptoms
40 " " " " "	mild "
60–80 " " " " "	severe "
100 " " " " "	coma, rigidity and paralysis

Presumably some such relationship could be demonstrated with other sulfonamides, although the blood level corresponding to a given toxic effect would naturally vary. In contrast with such data, the extreme lack of toxicity of one sulfonamide should be cited Welch, Mattis and Latven[8] showed that monkeys could be given one gram per kilogram body weight of sodium succinyl sulfathiazole, intravenously, daily for as long as ten days with negligible effects, although one kidney had been removed and although blood levels as high as 167 mg. per cent were encountered These same workers showed that in the mouse a dose of 7 grams per kilogram body weight of the sodium salt given intraperitoneally killed about 40 per cent of a group of mice. In such animals the blood level of the succinylsulfathiazole rose to 775 mg. per cent accompanied by 10 mg. per cent of sulfathiazole. In man the largest acute dose of sulfonamide that I have encountered is reported by Cutts and Bowman.[9] Nearly half a gram of sodium sulfapyridine per kilogram body weight was inadvertently administered to this patient over a period of 10 hours by intravenous injection. The principal symptoms were restlessness, irritability, gross hematuria and jaundice. The authors attributed the patient's complete recovery to his youth and good condition.

Other acute effects of a large dose of a sulfonamide may be a decreased clearance of creatinine by the kidneys and an acidosis with reduction of carbon dioxide combining power of the blood (e g , a fall from 56 volumes per cent to 39 volumes per cent was sometimes associated with a marked fall of blood pH to such low levels as 7.18 and 7.25). Litchfield[10] found no evidence of a local toxic effect of sulf-

[7] Klein, J. R., & Harris, J. S. Jour. Biol. Chem 124: 613. 1938.
[8] Welch, A. D., Mattis, P. A., & Latven, A. R. Jour. Pharmacol. 75: 231. 1942.
[9] Cutts, F. B., & Bowman, R. O. New Engl. Jour. Med 225: 448. 1941.
[10] Litchfield, J. T., Jr. Jour. Pharmacol. 67: 212. 1939.

anilamide on either the frog heart or the guinea-pig uterus when the perfusion fluid or bath contained 100 mg. per cent. On the other hand, sulfonamides definitely affect isolated cells growing as cultures, as demonstrated by the experiments of Jacoby, Medawar and Willmer.[11] These authors reported that the percentage of mitoses in cultures of fibroblasts or macrophages of fowl's blood was reduced in the presence of sulfonamides. The effect was reversible except in the presence of enormous concentrations of sulfanilamide. Sulfathiazole in terms of a saturated solution appeared to be more toxic than sulfapyridine or sulfadiazine.

The effects of repeated doses of sulfonamides over a period of days or weeks have far greater bearing on possible harmful effects in man than do the acute toxic experiments that have just been considered. One of the favorite methods of determining chronic toxic effects is to administer repeated doses to groups of growing rats and to determine the effects not only on growth and other phenomena but also on the organs of the animals sacrificed at the end of the experiment. The cumbersome method of attempting to administer repeated daily doses by injection or stomach tube has been largely replaced by a method advocated by Bieter and his associates.[12] These investigators recommend that the drug be mixed with the food at one or more levels so that a continuous intake be assured, provided that the animals eat the food. By the use of this method it is also possible with specially designed cages and food containers to estimate accurately the daily food consumption and hence the daily dose of drug ingested by mice or rats. The other laboratory animals commonly used for similar experiments in which, however, it is usually preferable to administer individual doses are rabbits, dogs and monkeys The last-named species is particularly useful since the signs and symptoms of poisoning presumably have more significance with regard to the use of such drugs in man. For example, in monkeys, it would be important not only to determine the effects on weight together with associated symptoms, but also effects on the morphology as well as on the glucose and carbon dioxide combining power of the blood and effects on the kidneys as demonstrated by possible changes in the composition of the urine. At death, or at the sacrifice of the animals, it would be of great interest to learn what gross and pathological changes in important organs like the liver, kidneys, spleen, and lungs are present.

A new aspect of chronic toxic effects of sulfonamides was first de-

[11] Jacoby, E., Medawar, P. B., & Willmer, E. N. Brit. Med. Jour. 2: 149. 1941.
[12] Bieter, R. N., Larson, W. P., Cranston, E. M., & Levine, M. J. Jour. Pharmacol. 66: 8. 1939.

scribed by Black, Overman, Elvehjem and Link.[13,14] These authors pointed out that a drug like sulfaguanadine, designed for intestinal bacteriostasis, interferes with the bacterial synthesis of vitamins or vitamin-like substances. Growing rats placed on a *synthetic* diet containing 0.5 per cent of sulfaguanadine exhibit in several weeks a hypoprothrombinemia as well as a marked reduction of the rate of growth The hypoprothrombinemia can be corrected by the administration of 2-methyl, 1,4 naptho-quinone. The growth deficiency can be corrected by the administration of para-aminobenzoic acid or of a suitable extract, which the authors conclude is potent because of its content of folic acid. According to Welch[15] a similar effect is produced if the diet contains 1 per cent of succinyl sulfathiazole, with the exception that liver extract but not para-aminobenzoic acid makes possible the resumption of growth. A still more recent report has been published by Daft, Ashburn and Sebrell.[16] These authors concluded that a 1 per cent concentration in the diet of either succinyl sulfathiazole or sulfaguanadine, if continued for a much longer period, leads to a variety of pathological findings in the blood and certain tissues, including voluntary muscle. A dermatitis appeared which could be corrected by the administration of biotin. References to the latest work in this field will be found in the communications of Wright and Welch[17] and by Gant and her colleagues.[18]

MacKenzie, MacKenzie and McCollum[19] pointed out that a marked thyroid hypertrophy with pronounced hyperplasia of the epithelium of the thyroid appears in rats receiving sulfaguanadine in the diet for a long period. Subsequent reports by Mackenzie and MacKenzie[20] and by Astwood and coworkers[21] confirmed the finding that iodides do not affect the progress of the hyperplasia. Since (1) the calorigenic action of thyroid extract or thyroxine is not impaired and (2) hypophysectomy prevents the thyroid hyperplasia and hypertrophy, it is concluded that the following chain of events occurs: the administered sulfonamide prevents formation of thyroid hormone; this deficiency leads to marked hypersecretion of anterior pituitary thyrotropic hormone which, although stimulating the thyroid gland morphologically, does not lead to the expected formation of thyroid hormone because

---

13 Black, S., McKibbin, J. M., & Elvehjem, C. A. Proc. Soc. Exp. Biol. Med. 47: 308. 1941
14 Black, S., Overman, R. S., Elvehjem, C. A., & Link, K. P. Jour. Biol. Chem. 145: 137. 1942.
15 Welch, A. D. Federation Proc. 1: 171. 1942.
16 Daft, F. S., Ashburn, L. L., & Sebrell, W. H. Science 96: 321. 1942.
17 Wright, L. D., & Welch, A. D. Science 97: 426. 1943.
18 Gant, O. K., Ransome, B., McCoy, E., & Elvehjem, C. A. Proc. Soc. Exp. Biol. Med 53: 276. 1943
19 MacKenzie, J. B., MacKenzie, C. G., & McCollum, E. V. Science 94: 518. 1941.
20 MacKenzie, C. G., & MacKenzie, J. B. Endocrinology 32: 185. 1943.
21 Astwood, E. B., Sullivan, J., Bissell, A., & Tyslowitz, R. Endocrinology 32: 210. 1943.

the thyroid is poisoned by sulfonamide. The phenomenon can be produced in mice, rats, and dogs but not in fowls and guinea pigs. It is much more evident after sulfapyridine or sulfadiazine than after sulfanilamide or sulfathiazole.

The "chronic" toxic effects of sulfonamides in man have sometimes appeared to be remarkably low. For example, Fitch[22] observed only inconsequential symptoms such as drug rash and drug fever in a ten-year-old child receiving 28 grams of sulfamethylthiazole and 203 grams of sulfathiazole in a period of three weeks. The remarkable report of Thomas, France, and Reichsman[23] deals with 55 patients who were adolescents or young adults for the most part. These patients who had a history of repeated attacks of rheumatic fever were given prophylactic doses of sulfanilamide over a period of four years. During 8 months of the first year each patient received about 215 grams of the drug. During 8 months of each of the three succeeding years the total dose was increased to 290 grams. Therefore, the total dose of each individual during a period of four years was approximately 1085 grams. There were remarkably few toxic symptoms. Occasionally there occurred a transient vertigo. The worst symptoms were described as drowsiness and itching. No granulocytopenia was observed, although the total leukocyte count tended to fall. On the other hand Sutliff, Helpern, Griffin and Brown[24] attempted to estimate the extent to which sulfonamides could be held responsible for deaths in New York City in 1941. They concluded that one death in about 2600 could be attributed to sulfonamides. They believed such drugs to be responsible for a higher proportion of deaths in patients ill with pneumonia (one in 1600).

A number of authors and their associates have published tables outlining the toxic effects of various sulfonamides (Long, Finland, Flippin, Spink, and others). The table reproduced as TABLE 2 was published by Dowling and Lepper.[25] In this table the comparative toxic effects of sulapyridine, sulfathiazole and sulfadiazine as observed by two investigators suggest the type and frequency of toxic effects that may be expected in man. Naturally the percentage of toxic complications observed will vary from clinic to clinic depending upon ade-

[22] Fitch, T. S. P. Arch. Pediat. 57: 119. 1940.
[23] Thomas, C. B., France, R., & Reichsman, F. Jour. Am Med. Assoc 116: 551. 1941.
[24] Sutliff, W. D., Helpern, M., Griffin, G., & Brown, H. Jour. Am. Med. Assoc. 121: 307. 1943.
[25] Dowling, H. F., & Lepper, M. H. Jour. Am. Med. Assoc. 121: 1190. 1943.

quacy of observation, technique of dosage, diseases treated, age of patients, etc. For the purpose of this presentation, however, the table of Dowling and Lepper furnishes a satisfactory basis for subsequent discussion.

TABLE 2

TOXIC EFFECTS OF SULFAPYRIDINE, SULFATHIAZOLE AND SULFADIAZINE IN MAN (Dowling and Lepper[25])

Toxic Reaction	Sulfapyridine No. of patients	Sulfapyridine Per cent	Sulfathiazole No. of patients	Sulfathiazole Per cent	Sulfadiazine No. of patients	Sulfadiazine Per cent
Vomiting	102	20.1	19	5.9	9	1.4
Renal calculus	8	1.6	9	2.8	10	1.5
Drug fever, dermatitis and/or conjunctivitis	12	2.4	20	6.2	13	2.0
Mental confusion	12	2.4	5	1.6	10	1.5
Leukopenia (with or without granulocytopenia)	6	1.2	4	1.2	6	0.9
Acute hemolytic anemia	7	1.4	1	0.3	1	0.2
Leukocytosis	0	0	0	0	2	0.3
Yellow vision	0	0	1	0.3	1	0.2
Peripheral neuritis	1	0.2	0	0	0	0
Total patients with toxic reactions	149	29.4	58	11.8	51	7.7
Patients with no toxic reactions	359	70.6	263	88.2	609	92.3
Total patients treated	508	100.0	321	100.0	660	100.0

CHANGES IN THE BLOOD

Granulocytopenia

Leukopenia is frequently observed following the administration of sulfonamides to patients in whom there is no pre-existing leukocytosis. Leukopenia may herald one of the most serious toxic complications of sulfonamide therapy—agranulocytosis. In many instances, there is found simply a reduction in the number of granulocytes in the blood (granulocytopenia). If this is observed the sulfonamide should be withdrawn unless compelling reasons urge a continued cautious administration. If granulocytopenia is detected during sulfonamide therapy and the drug is continued, there is real likelihood that the granulocytes will disappear from the blood entirely and that fever associated with marked inflammation and necrotic changes in the pharynx will be observed. A number of authors have reported that withdrawal of the sulfonamide when granulocytopenia is detected is no guarantee that

agranulocytosis will not appear. It is impossible to state how many individuals have died of agranulocytosis caused by sulfonamide administration. However, there probably have been hundreds of fatalities. Usually granulocytopenia or its serious terminal development, agranulocytosis, is observed in patients who receive large total doses of sulfonamides over periods of 18–25 days. This complication rarely appears earlier than about 12 days after the beginning of therapy. In the first two cases reported in which sulfanilamide was the offending drug, 54 grams of sulfanilamide were administered to one patient during 19 days[26]; to the other patient 44 grams of drug were administered over 25 days.[27] Young's hematological data are presented in TABLE 3

TABLE 3

AGRANULOCYTOSIS (disappearance of "polymorphs" and other leukocytes with granules) FOLLOWING THE ADMINISTRATION OF SULFANILAMIDE*

Day	3	5	9	12	18	23	24†
Hemoglobin (per cent)	90						82
Leukocytes (per cmm)	12,000	9,800	8,200	9,700	7,800	1,800	2,300
Polymorphs (per cmm)	6,600	5,047	4,633	6,984	5,694	0	0
Lymphocytes " "	4,344	3,577	2,214	2,085	1,755	1,800	2,300
Monocytes " "	660	784	861	388	351	0	0
Eosinophils " "	396	392	410	194	0	0	0
Basophils " "	0	0	41	0	0	0	0
Myelocytes " "	0	0	41	49	0	0	0
Arneth count *I* per cent	32	32	41	59	58	0	0
II " "	45	44	41	34	34	0	0
III " "	21	24	17	7	8	0	0
IV " "	2	0	1	0	0	0	0
V " "	0	0	0	0	0	0	0
I + *II* (per cmm)	5,082	3,800	3,772	6,510	5,244	0	0
Lobes per 100 polymorphs	193	192	178	148	150	0	0
Toxic granulation in polymorphs per cent	0	0	0	0	2	—	—

* The results of repeated examination of the blood of a patient receiving 3 gm of sulfanilamide daily on days 1-19 are given above The data were published by Young [26]

† All counts 3 hours after death

Any sulfonamide may provoke granulocytopenia and agranulocytosis However, the appearance of this syndrome has occurred more frequently after the administration of sulfanilamide, perhaps because the total number of cases treated by sulfanilamide is greater than that of any other sulfonamide

Agranulocytosis has sometimes been likened to pernicious anaemia and has been called pernicious leukopenia The etiology so far as sulfonamides are concerned, probably, is a depressant or destructive

[26] Young, C. J. Brit. Med. Jour. 2: 105. 1937.
[27] Plumer, H. H. New Engl. Jour. Med. 216: 711 1937

effect of the responsible drug on the cells of the bone marrow from which granulocytes are derived. In agranulocytosis there may occur a complete myeloid aplasia.

Effects on Erythrocytes or Hemoglobin

It is more convenient to continue the discussion of the effects of sulfonamides on other aspects of the morphology or chemistry of the blood than to consider other drug complications in the exact order of their importance. The second effect on the blood that immediately comes to mind is hemolytic anemia.

Hemolytic anemia is characterized by an increase of the fragility of the erythrocytes accompanied by intravascular hemolysis, anemia, and a compensatory increase in the proportion of young erythrocytes or reticulocytes. Usually hemolytic anemia, owing to sulfonamide therapy, occurs during the first eight days of therapy and is observed most frequently about the fifth day. A drug like sulfanilamide probably causes this change more frequently in children than in adults. Anemia is observed because the compensatory production of reticulocytes does not keep pace with the destruction of mature erythrocytes. In severe cases, 30 per cent or more of the circulating red cells may undergo destruction representing a loss of 500–700 grams of pigment destroyed or excreted as hemoglobin or derivatives of hemoglobin. Hemolytic anemia may be caused by sulfanilamide, sulfapyridine, sulfathiazole, sulfadiazine or diaminodiphenylsulfone. (Another type of anemia in which the poisonous effect of the drug is exerted on the blood-forming cells of the bone marrow, aplastic anemia, has been observed only rarely in association with sulfonamide therapy.)

There has been considerable discussion of the mechanism by which sulfonamides increase the fragility of the erythrocytes. Perhaps as satisfactory a hypothesis as any is that oxidation products of sulfonamides are the responsible agents. For example, Emerson, Ham and Castle[28] pointed out that para-aminophenol and phenylhydroxylamine, perhaps produced in the body after the administration of a sulfonamide, readily cause hemolysis both *in vitro* and *in vivo*. On the other hand Thorpe, Williams and Shelswell,[29, 30] contradicting James,[31] could obtain no evidence that these or related substances occurred in the urine, although 6–12 per cent of the dose of sulfanilamide appeared as a phenol-hydroxy body conjugated with sulfate. Rimington and Hem-

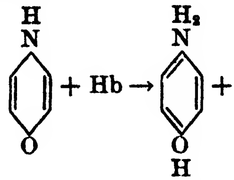
[28] **Emerson, C. P., Ham, T. H., & Castle, W. B.** Jour. Clin. Investigation 20: 451. 1941.
[29] **Thorpe, W. V., Williams, R. T., & Shelswell, J.** Biochem. Jour. 35: 52. 1941.
[30] **Shelswell, J., & Williams, R. T.** Biochem. Jour. 34: 528. 1940.
[31] **James, G. V.** Biochem. Jour. 34: 640. 1940.

mings[32, 33] attributed to an obscure toxic effect of a sulfonamide, like sulfanilamide, the abnormally high urinary and fecal excretion of porphyrins. The effect persists several weeks after cessation of therapy. It is of interest that certain coproporphyrins, hematin derivatives containing no iron, vary in their excretion ratios in different animals. In the rat receiving sulfanilamide, coproporphyrin III is the chief excretion product in this category, whereas, in man, after similar treatment about equal amounts of coproporphyrins I and III are excreted. Rimington and Hemmings[33] concluded that a substance containing an unsubstituted or a potentially free aromatic amino group causes an abnormal increase in porphyrin excretion. They believed that such an effect is more certain if the drug in question undergoes degradation or oxidation to hydroxylamine or imino-quinone. Usually the excretion of abnormal amounts of porphyrins is associated with methemoglobinemia.

In addition to effects on erythrocyte fragility, sulfonamides also bring about changes in the blood pigments. In the preceding paragraph, methemoglobinemia was mentioned as a common finding associated with altered porphyrin excretion. Cyanosis occurs very frequently in patients receiving sulfanilamide and there has been considerable debate concerning its explanation. Some authors believe that cyanosis is not necessarily associated with a change in hemoglobin and they have suggested that a colored derivative of the drug is responsible for the cyanosis.[34, 35, 36] Indeed, James reported that such a pigment can be produced *in vitro* by the action of light in the presence of iron and an oxidizing agent such as H_2O_2. Others believe that methemoglobin is principally responsible for the cyanosis and that the failure of some authors to recognize its presence is to be attributed to technical deficiencies. Harris and Michel[37] reported that 58 per cent of patients receiving sulfanilamide exhibited methemoglobinemia. The functional hemoglobin may be reduced 15–30 per cent, but this reduction is not necessarily related to the blood level of a drug like sulfanilamide.[38] Sulfonamides can produce methemoglobinemia in the fowl, the pigeon, and rat.[35, 39] Rimington and Hemmings[33] have suggested that a hypothetical oxidation product of sulfonamide causes the conversion of hem-

32 Rimington, C., & Hemmings, A. W. Lancet. 1: 770. 1938.
33 Rimington, C., & Hemmings, A. W. Biochem. Jour 33: 960. 1939.
34 Marshall, E. K., Jr., & Walzl, E. M. Bull. Johns Hopkins Hosp. 61: 140. 1937.
35 Webb, J. T., & Kniazuk, M. Jour. Biol. Chem. 128: 511. 1939.
36 James, G. V. Biochem. Jour. 34: 636. 1940.
37 Harris, J. S., & Michel, H. O. Jour. Clin. Investigation 18: 507. 1939.
38 Wendel, W. B. Jour. Clin. Investigation 18: 179. 1939
39 Richardson, A. P. Bull. Johns Hopkins Hosp. 65: 446. 1939; Jour. Pharmacol. 72: 99. 1941.

oglobin into methemoglobin, perhaps according to the following scheme:



+ met Hb (ortho- instead of para-imino-quinone might be the oxidant)

As first discovered by Wendel,[40] methylene blue effectively reduces the methemoglobin caused by the administration of sulfanilamide. Intravenous doses of the order of 0.1–1 mg. per kg may reduce nearly all the circulating methemoglobin in about 45 minutes. Oral doses of the order of 1 gram likewise are highly effective although the rate of reduction is naturally slower. Wendel[38] suggests that methylene blue is reduced in the body to the leuko form and that the latter reduces methemoglobin to hemoglobin.

Another blood pigment to which considerable attention has been given is sulfhemoglobin following the first report of Colebrook and Kenney[41] in 1936. The existence of sulfhemoglobinemia is illustrated by the report of Harris and Michel,[37] who found the pigment in 8 per cent of patients receiving sulfanilamide. Richardson[19] concluded that this pigment instead of methemoglobin is found in the blood of mice receiving either sulfanilamide or sulfapyridine. Sulfathiazole did not produce sulfhemoglobinemia in mice. On *a priori* grounds and on the basis of Richardson's work, the belief of physicians that the concurrent administration of sulfates and sulfonamides must be avoided to prevent possible sulfhemoglobinemia appears to have no foundation. When investigators refer to sulfhemoglobin they usually refer to an abnormal pigment, not methemoglobin, with an absorption band at about 620 millimicrons. There is no good evidence that such a pigment is identical with that derived from hemoglobin by the action of H_2O_2 on hemoglobin in the presence of inorganic sulfide ion.[42] (Harris and Michel suggested that an active derivative of sulfonamide in association with sulfide acts upon hemoglobin to produce sulfhemoglobin.) The latest view of Fox and Ottenberg[43] is that what has been identified as sulfhemoglobin in patients receiving sulfanilamide is really the methemalbumin of Fairley.[44]

---

[40] Wendel, W. B. Jour. Amer. Med. Assoc. 109: 1216. 1937
[41] Colebrook, L., & Kenney, M. Lancet 1: 1279. 1936.
[42] Sunderman, F. W. Personal communication.
[43] Fox, C. L., Jr., & Ottenberg, R. Jour. Clin. Investigation 20: 593. 1941.
[44] Fairley, N. H. Nature 142: 1156. 1938.

Two other changes in the blood deserve mention. Thrombocytopenia with purpura is a rare complication of sulfonamide therapy. In about one-half the reported cases death has occurred. After Southworth[45] recognized that the clinical use of sulfanilamide is associated with a reduced CO_2 combining capacity of the blood ("acidosis"), a number of confirmatory reports have appeared Beckman, Krayer, and Bauer[46] attributed this change to a renal loss of bicarbonate and sodium with retention of chloride owing to deficient tubular reabsorption of sodium and bicarbonate. The explanation of the exact mechanism of the acidosis rests upon the discovery of Mann and Keilin[47] that sulfanilamide and other substances of the sulfonamide group poison the carbonic anhydrase of Meldrum and Roughton. This enzyme, present in high concentration in erythrocytes, catalyzes the release of carbon dioxide from carbonic acid and bicarbonate. Mann and Keilin showed that as little as 2×10^{-6}M concentration of sulfanilamide poisons the enzyme, whereas sulfonamides without a free sulfonamide group (e g. sulfapyridine, sulfathiazole, or sulfadiazine) are without action Subsequent work in man by Roughton, Dill and their coworkers,[48, 49] as well as by Wood and Favour,[50] fully confirm the belief that sulfanilamide, but not sulfathiazole or sulfadiazine, produces a fall of the carbon dioxide combining power of the blood Apparently sulfanilamide poisons the carbonic anhydrase so that the release of carbon dioxide in the lungs is extremely deficient. Carbonic anhydrase is also found in high concentration in (acid) secreting cells of the gastric mucosa and in the cortex of the kidney where it appears to catalyze the reabsorption of bicarbonate Therefore, if the renal enzyme be poisoned, bicarbonate is lost and the alkalinity of the urine is increased[51, 52] An explanation of the results of Beckman, Krayer, and Bauer[46] is thus furnished

EFFECTS OF SULFONAMIDES ON THE KIDNEYS

Only after N^1-heterocyclic derivatives of sulfanilamide were introduced was there recognition of serious toxic effects of sulfonamides on the kidneys Because these new derivatives are the drugs of choice in the treatment of bacterial infections, renal complications attending

[45] Southworth, H. Proc Soc. Exp Biol Med 36: 58 1937
[46] Beckman, W., Krayer, O., & Bauer, W. Jour Clin Investigation 20: 455 1941.
[47] Mann, T., & Keilin, D. Nature 146: 164 1940
[48] Roughton, F. J. W., Dill, D. B., Darling, R. C., Graybiel, A, Knehr, C. A, & Talbott, J. H. Am. Jour. Physiol 135. 77. 1941.
[49] Roughton, F. J. W., Darling, R. C., Forbes, W. H, Horvath, S. M., Robinson, S., & Talbott, J. H Am. Jour Physiol 137: 598. 1942
[50] Wood, W B., & Favour, C. B. Jour Clin Investigation 20: 433. 1941
[51] Davenport, H. W., & Wilhelmi, A. E. Proc Soc. Exp Biol Med. 48: 53. 1941.
[52] Höber, R Proc. Soc Exp. Biol Med 49: 87 1942

therapy by such drugs deserve the most careful scrutiny. Up to the present, sulfapyridine and sulfathiazole have frequently been responsible for renal damage. Sulfadiazine has been implicated much less frequently. Usually the damage is the result of mechanical obstruction which may be as low as the ureters or as high as the nephron, although, in animals, there may appear focal injury of the tubules and glomeruli independent of any precipitation of free or acetylated drug in the tubules and pelves of the kidneys (e g. sulfathiazole).[53] If it is agreed that mechanical obstruction is the important factor of renal damage, then the following factors determine the extent to which a given drug will produce renal pathological changes:

1. The solubility of free and acetylated drug is of great importance FIGURE 3 illustrates the relationship with reference to sulfathiazole as

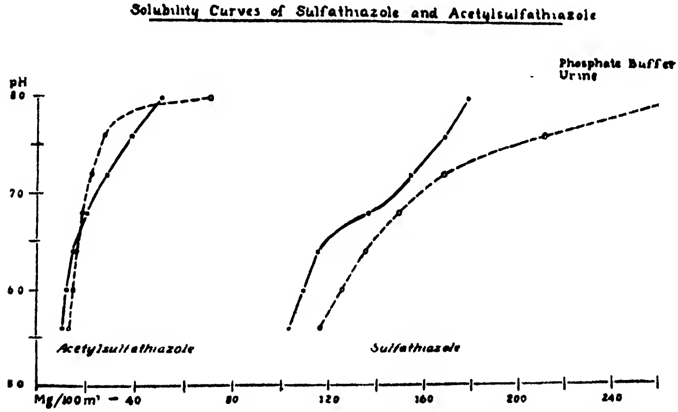


FIGURE 3 Solubility of sulfathiazole and acetylsulfathiazole (from Sunderman Pepper and Benditt[54])

reported by Sunderman, Pepper and Benditt[54] It is apparent that after acetylation the solubility of sulfathiazole is greatly reduced and the danger of its precipitation in the urine is correspondingly increased An experiment comparing the solubility of free and acetylated sulfadiazine and sulfamerazine (2-sulfanilamide-4-methylpyrimidine) is graphically portrayed in FIGURE 4 reproduced from the paper of Welch, Mattis, Latven, Benson, and Shiels[55] It is of great interest that the

[53] Rake, G, van Dyke, H. B., & Corwin, W C. Am Jour Med Sci 200 358 1940
[54] Sunderman, F W, Pepper, D. S., & Benditt, E. Am Jour Med Sci 200: 790 1940
[55] Welch, A. D., Mattis, P A., Latven, A. R, Benson, W. M, & Shiels, E H Jour. Pharmacol 77: 357. 1943.

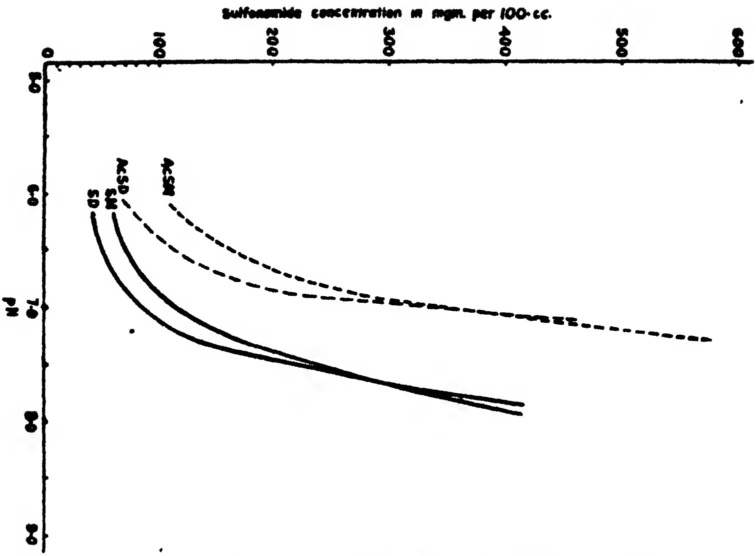


FIGURE 4 Solubility of sulfadiazine (SD), sulfamerazine (2-sulfanilamide-4-methylpyrimidine, SM) and of their acetyl derivatives (AcSD, AcSM) in urine at 37.5° C (from Welch, Mattis, Latven, Benson and Shiels[55]).

acetylated compounds are more soluble than the free drugs and that the solubility of acetyl sulfamerazine is great enough so that the danger of urinary precipitation is very small in comparison with the acetyl derivative of a drug like sulfathiazole. (Rose, Martin, and Bevan[56] reported that sulfamethazine (2-sulfanilamide-4, 6-dimethylpyrimidine) is more soluble than the acetylated derivative; but this relationship has not been confirmed by others.)

2 A second factor of real significance is the rate of clearance of the free or acetylated drug during urinary excretion. If the drug in question is completely or nearly completely cleared, the ultrafiltrate of plasma passing from the glomerulus to the tubular lumen contains a higher and higher concentration of drug as the fluid flows to the collecting tubule, since water and useful ions, including bicarbonate, are reabsorbed by the tubular epithelium, whereas the drug is poorly reabsorbed. Thus not only does the concentration of drug in the tubular fluid rapidly rise but at least one other condition favoring its precipitation, lowering of pH, also appears. In the case of sulfathiazole it is

[56] Rose, F. L., Martin, A. R., & Bevan, H. G. L. Jour. Pharmacol. 77: 127. 1943.

probable that the acetylated derivative is more rapidly cleared than the free drug, which itself is noted for its rapid renal elimination.

3. The hydrogen ion concentration of the urine has just been mentioned as an important factor determining whether or not free or acetylated drug or both will be precipitated in the urine. The higher the pH of the urine the less likely is the formation of drug uroliths either microscopically or grossly.

In man, there is no necessary relation between total dose and urolith formation which has occurred following the administration of as little as 5 to 6 grams of sulfonamide or as much as 72 grams. The first clinical report suggesting urolith formation was that of Adalja.[57] Credit for the unquestioned recognition of this complication belongs to Antopol and Robinson[58] and to Gross, Cooper, and Lewis,[59] who simultaneously reported experiments on rats receiving sulfapyridine. In man, the complication of drug precipitation in the urinary tract has occurred following the administration of sulfapyridine, sulfathiazole, sulfadiazine, and sulfaguanidine, ranging from transient hematuria to such serious symptoms as oliguria, renal colic, anuria and even death. Probably sulfathiazole is the worst offender among drugs in common use today. Some idea of the pathological complications as they are observed in the rat and monkey after repeated administration of sulfapyridine or sulfathiazole may be gained from FIGURES 5–12.[53]

Discussion of this topic would not be complete without mention of the remarkable properties of succinylsulfathiazole as reported by Welch, Mattis, and Latven[8] in 1942. In contrast to sulfathiazole, its N^4-succinyl derivative, although producing a severe crystalluria as after the intravenous injection daily of one gram of the sodium salt per kilogram body weight into unilaterally nephrectomized monkeys, causes neither significant renal damage nor the formation of any drug concretions. Succinylsulfathiazole may appear in the urine in the remarkably high concentration of 12 grams per cent.

---

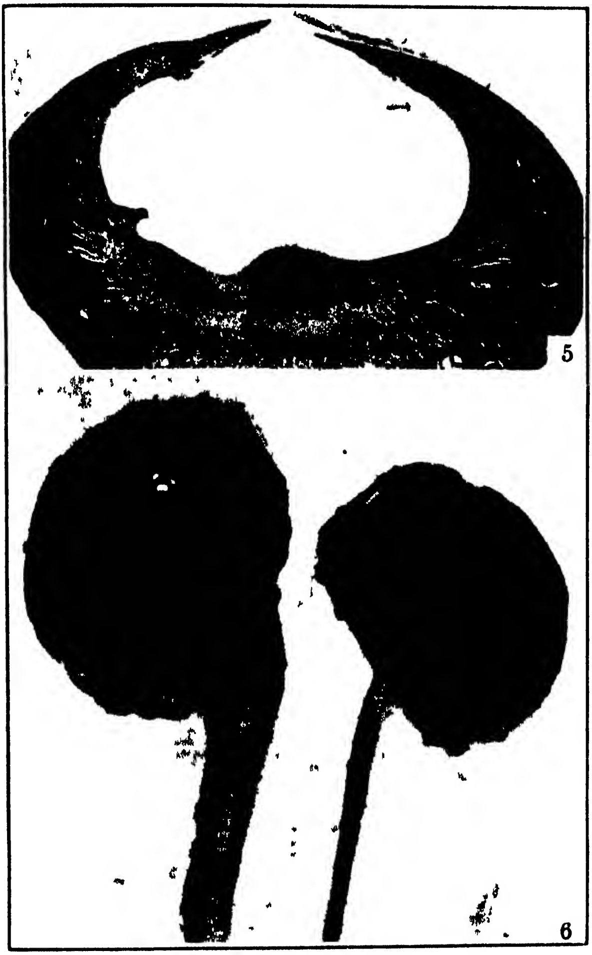
[57] **Adalja, K. V.** Brit. Med. Jour. 1: 643. 1939.
[58] **Antopol, W., & Robinson, H.** Proc. Soc. Exp. Biol. 40: 428. 1939
[59] **Gross, P., Cooper, F. B., & Lewis, M.** Proc. Soc. Exp Biol & Med 40: 448. 1939

PLATE 1

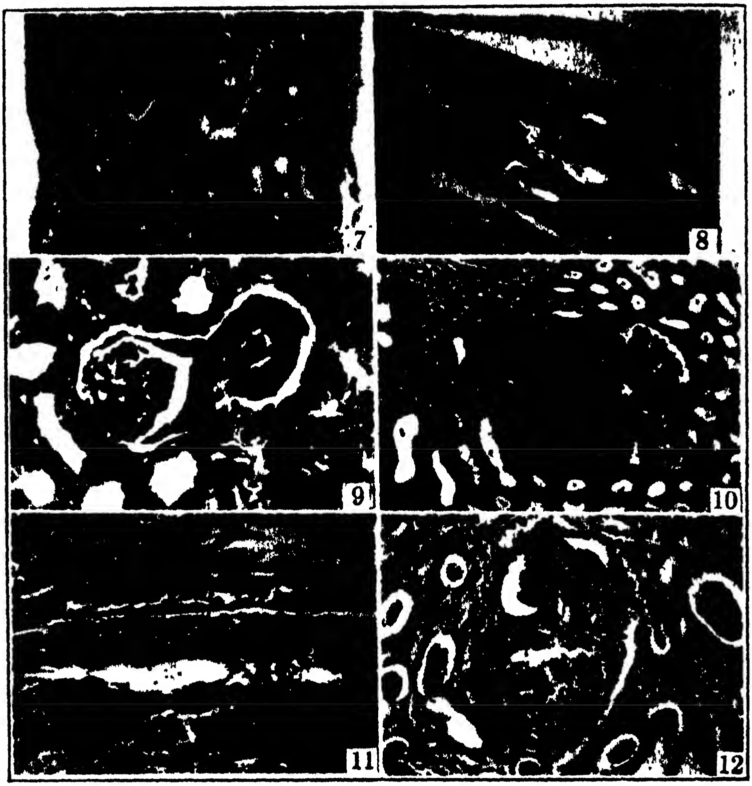
FIGURE 5 Sulfapyridine Rat kidney showing hydronephrosis with compression of the organ and dilatation of tubules in cortex and medulla (X7 H and E)

FIGURE 6 Sulfapyridine Monkey 49 Kidney Hydronephrosis and hydroureter on the right side

(Figures 5 and 6 from Rake, van Dyke, and Corwin [82])



VAN DYKE: TOXIC EFFECTS OF SULFONAMIDES



VAN DYKE. TOXIC EFFECTS OF SULFONAMIDES

PLATE 2

FIGURE 7. Sulfapyridine. Rat kidney showing the total width of the organ. Some tubules dilated, others collapsed and surrounded with scar tissue (X. 110. H. and E.).

FIGURE 8 Sulfapyridine. Rat kidney showing the outline of crystals in a tubule and surrounding leukocytic infiltration (X 110. H. and E.).

FIGURE 9 Sulfapyridine. Monkey kidney showing albuminous cast filling the glomerular capsule and commencement of the tubule (X 260. H. and E.).

FIGURE 10 Sulfapyridine Monkey kidney showing leukocytic infiltration in and around tubules forming an abscess (X 110. H. and E.)

FIGURE 11 Sulfathiazole. Monkey kidney showing the outline of crystals in tubules and surrounding leukocytic infiltration (X 110. H. and E).

FIGURE 12. Sulfathiazole. Monkey kidney showing leukocytes in and around the tubules (X 110. H. and E.).

(Figures 7–12 from Rake, van Dyke, and Corwin.[53])

PLATE 3

FIGURE 13 Morbilliform rash with definite hemorrhagic lesions following the administration of sulfanilamide (figure from Hageman and Blake[illegible])

FIGURE 14 Nodular rash following the administration of sulfathiazole (figure from Volini, Levitt and O'Neill[illegible]).

FIGURE 13

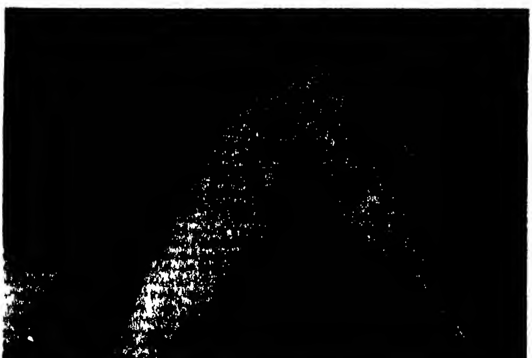


FIGURE 14

EFFECTS OF SULFONAMIDES ON NERVOUS TISSUE

Effects of sulfonamides on the central nervous system were first observed in animals during the determination of acute toxic effects of single doses. Under these conditions the changes produced are entirely in the central nervous system and consist of purposeless running movements, spastic paralysis, convulsions, coma and a variety of intermediate symptoms, all of which are not peculiar to a particular sulfonamide. Deleterious cerebral effects of ordinary therapeutic doses of sulfanilamide are well recognized and have led to the recommendation that for 3 to 7 days after cessation of sulfanilamide therapy the individual should not undertake any responsible tasks such as piloting an airplane, operating an automobile or making important decisions.

For the experimental determination of delayed or chronic toxic effects of sulfonamides on the nervous system, birds (pigeons, fowls) appear to be the best experimental animals, since they are much more sensitive to nervous injury, especially of the peripheral nerves, than are mammals. Earlier work was done by Hüllstrung and Krause,[60] who concluded that the mono- and di-methyl derivatives of sulfanilylsulfanilamide can cause polyneuritis in pigeons. Rosenthal[61] and Nelson[62] in the following year were not able to demonstrate that sulfanilamide had much effect on the central or peripheral nervous systems of fowls. In 1941 Bieter and his colleagues[63] studied the effects of six sulfonamides on the brain, spinal cord (including peripheral axones and myelin sheaths) and peripheral nerves of white Leghorn fowls. Substances which had the greatest toxic effect were sulfaphenylthiazole, dimethylsulfanilyl-sulfanilamide and sulfamethylthiazole. The three other sulfonamides listed in decreasing order of toxicity were sulfathiazole, sulfapyridine and sulfanilamide. Sciatic nerves of experimental fowls often contained a concentration of sulfonamide high in comparison with the blood, liver, brain and spinal cord.

In man peripheral neuritis has appeared in about 3 per cent of patients treated either with uliron (sulfanilyl-dimethylsulfanilamide) or sulfamethylthiazole. Cases have also been reported following the administration of sulfapyridine, sulfathiazole, sulfadiazine and sulfanilamide.[64] Sulfanilamide was reported by Bucy[65] to have caused neuritis of the optic nerve. Effects on the peripheral nerves are probably less

[60] Hüllstrung, H., & Krause, F. Deut. med. Wochschr. 64: 114. 1938.
[61] Rosenthal, S. M. Public Health Reports 54: 95. 1939.
[62] Nelson, A. A. Public Health Reports 54: 106. 1939.
[63] Bieter, R. N., Baker, A. B., Beaton, J. G., Shaffer, J. M., Seery, T. M., & Orr, B. A. Jour. Amer. Med. Assoc. 116: 2231. 1941.
[64] Little, S. C. Jour. Amer. Med. Assoc. 119: 467. 1942.
[65] Bucy, P. C. Jour. Amer. Med. Assoc. 109: 1007. 1937.

specific than those on the spinal cord or brain. Sulfonamides have also caused toxic psychosis and even encephalomyelitis. A peculiarity of sulfathiazole is that it may cause postoperative epileptic seizures if it is applied to the surface of the brain or at the closure of a craniotomy wound. Such an effect is not produced by sulfanilamide, sulfapyridine, sulfadiazine or sulfacetamide.[66, 67]

CUTANEOUS EFFECTS OF SULFONAMIDE THERAPY

Effects of sulfonamides on the skin have attracted considerable attention, since these cutaneous manifestations may occur frequently. Usually they do not put in appearance until after a lapse of 10–12 days' treatment. The frequency in patients is illustrated by the following percentages: sulfanilamide, 3–10 per cent; sulfapyridine, 10 per cent; sulfathiazole, 13 per cent or more. The types of cutaneous changes are varied. Morbilliform rashes are frequent and were described in 1937 by Hageman and Blake[68] (FIGURE 13). Scarlatiniform and urticarial rashes have been described If the toxic effect is of great severity otherwise, there may be an accompanying exfoliative dermatitis. Specific rashes have been described only after the administration of sulfathiazole (see FIGURE 14, from the paper of Volini, Levitt, and O'Neill).[69] After the administration of sulfathiazole, conjunctivitis, with or without nodular cutaneous lesions resembling those of erythema nodosum, may appear. Schnee[70] has observed membranous inflammation of the conjunctiva and of the mucous membrane of the nose, mouth, pharynx and larynx following sulfathiazole therapy.

Considerable attention has been directed to the mechanism by which sulfonamides produce cutaneous changes. Drug idiosyncrasy, congenital or induced by single or repeated courses of sulfonamide therapy, has been one classification favored by some authors. Idiosyncrasy can, of course, appear months after a previous course and apparently can be induced by sulfanilamide, sulfathiazole or sulfadiazine. Sometimes photosensitization appears to be an important factor in determining whether or not a sulfonamide will induce inflammatory changes in the skin. Erskine[71] reported that ultraviolet irradiation in association with sulfonamide may provoke cutaneous inflammatory changes. Working in the Near East, Park and Platts[72] reported that

[66] Watt, A. C., & Alexander, G. L. Lancet 1: 493. 1942.
[67] Pilcher, C., Angelucci, R., & Meacham, W. F. Jour. Amer. Med. Assoc. 119: 927. 1942.
[68] Hageman, P. O., & Blake, F. G. Jour. Amer. Med. Assoc. 109: 642. 1937.
[69] Volini, I. F., Levitt, R. O., & O'Neill, H. E. Jour. Amer. Med. Assoc. 116: 938. 1941.
[70] Schnee, I. M. Brit. Med. Jour. 1: 506. 1943.
[71] Erskine, D. Brit. Med. Jour. 2: 104. 1939.
[72] Park, R. G., & Platts, W. M. Brit. Med. Jour. 2: 966. 1942.

toxic manifestations in the skin following sulfonamides were often related to the degree of exposure to sunlight. The skin of subjects containing large amounts of melanin, as in Maoris, was never affected by the administration of sulfonamides. A third explanation which has been offered is that cutaneous changes are an allergic response with or without an associated fever. In no instance has any investigator been able to show that the drug itself is responsible, so far as this can be demonstrated by patch tests, by intradermal injections or by a precipitin reaction between sulfonamide and the serum of individuals who are hypersensitive. Drug alone can neither produce passive sensitization in man or guinea pig nor anaphylaxis in the guinea pig

It is probable that a product of drug metabolism is really responsible for many of the cutaneous changes following sulfonamide therapy In this connection the experiments of Epstein[73] may be cited, as this author found that the intracutaneous injection of sulfanilamide, accompanied by ultraviolet irradiation of the area, might be followed by erythema and urticaria-like changes in 10 days When the injection was repeated later the reaction was induced in only 24 hours The author believed that local sensitization to a product of the irradiated intracutaneous sulfanilamide had occurred. Among metabolic products which have been implicated are certain hypothetical oxidation products such as phenylhydroxylamine and para-aminophenol Previous mention of the possibility of the appearance of such substances was made in connection with the discussion of the work of Rimington and Hemmings.[33] A remarkable case having a bearing on this question is that of Rogers [74] The patient in question had been sensitized to procaine and procaine derivatives since 1923. In 1938, following the administration of 1 6 grams of sulfanilamide, there appeared, after 36 hours, and continued during the next 3–10 days, erythema, itching, swelling and tenderness at all sites of operation into which procaine or another derivative or para-aminobenzoic acid had been injected to produce local anesthesia The symptoms spread from these primary sites so that maximum changes were present at about the third week. Although Rogers found that the patient exhibited a positive scratch test to procaine, this test was negative when sulfanilamide was similarly used. Rimington and Hemmings[33] believed that very useful information might have been obtained if Rogers had also employed for such tests phenylhydroxylamine or para-aminophenol.

Work in another direction has been reported by Abernathy, Bukantz,

---

[73] Epstein, S. Jour. Investig. Dermatol. 2: 43. 1939.
[74] Rogers, E. B. Jour. Amer. Med. Assoc. 111: 2290. 1938.

and Minor[75] and Wedum.[76] The first-named authors found that if diazotized sulfathiazole was coupled with globulin or serum albumin, this azoantigen was precipitated by the serum of a patient sensitized to sulfathiazole. Wedum made azoproteins of sulfanilamide, sulfapyridine and sulfathiazole. With such azoproteins he could produce anaphylactic shock, precipitin reactions and cutaneous sensitization. The azoproteins cross-reacted except perhaps in the production of anaphylactic shock. No reaction by any test with drug alone could be demonstrated. Para-aminobenzoic acid when coupled with proteins proved to be only a weak antigen.

It appears certain that para-aminobenzoic acid does not antagonize the toxic effects of sulfonamides. Strauss, Lowell and Finland[77, 78] showed that as much as 29 grams of para-aminobenzoic acid, administered as one gram every 2 hours, did not affect either fever or rash owing to sulfathiazole sensitivity

DRUG FEVER

Fever itself may appear without an accompanying rash (FIGURE 15), and may or may not be associated with leukocytosis Drug fever is, of course, a frequent complication of sulfonamide therapy and may be the outstanding manifestation of sensitization to one or more sulfonamides.

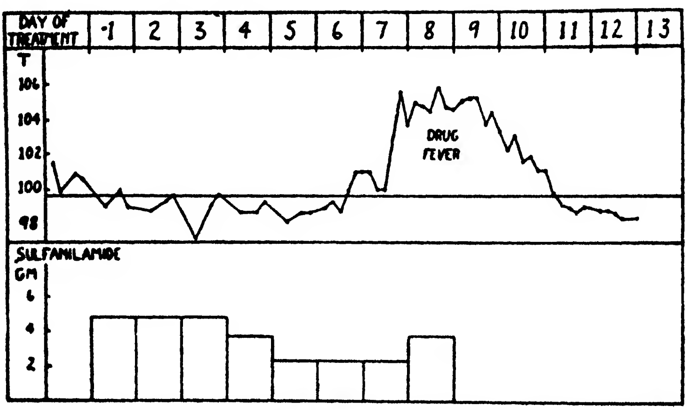


FIGURE 15 Drug fever caused by administration of sulfanilamide to a patient with pansinusitis (Hageman and Blake[59])

[75] Abernethy, T. J., Bukantz, S. C., & Minor, J. Jour Clin. Investigation 20: 433. 1941.
[76] Wedum, A. G. Jour Infectious Diseases 70: 173. 1942
[77] Strauss, E., Lowell, F. C., & Finland, M. Jour. Clin. Investigation 20: 189. 1941.
[78] Strauss, E., & Finland, M. Amer. Jour. Med. Sci. 201: 730. 1941.

Lyons and Balberor[79] found that drug fever appeared in one-third of a group of patients to whom sulfathiazole was administered a second time, although the first course of treatment caused no febrile reaction. Acquired sensitization to sulfathiazole characterized by a striking febrile reaction is shown in FIGURE 16, as reported by Nelson.[80]

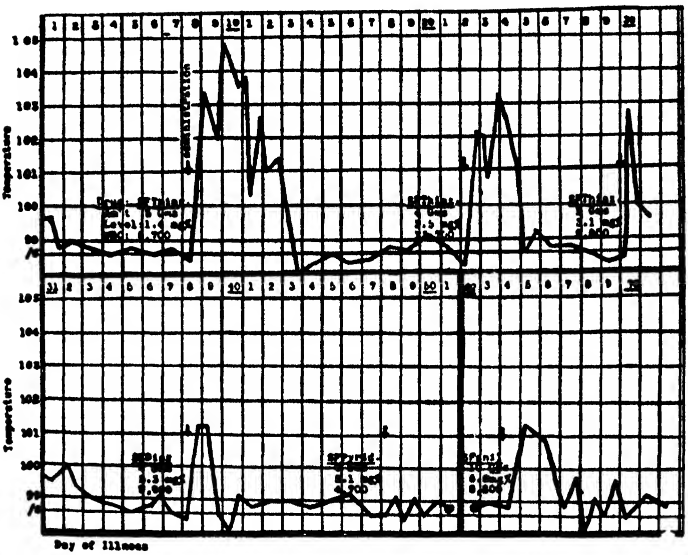


FIGURE 16 Acquired sensitivity to sulfonamides Seven months previously the patient had received 1 gm of sulfathiazole every four hours for three days The sulfonamide doses listed in this figure represent the sums of divided doses Other symptoms were nausea and headache with or without rash This figure was originally published by Nelson [80]

EFFECTS OF SULFONAMIDES ON THE LIVER

Fortunately, severe liver damage is not a frequent manifestation of poisoning by sulfonamides. Hemolytic jaundice is perhaps the least serious hepatic disturbance caused by these drugs. A toxic hepatitis may be associated with exfoliative dermatitis and indicates a severe general toxic effect of whatever sulfonamide may have been employed The most serious effect on the liver is the production of an acute yellow atrophy. This change has been reported in patients who have received 27–45 grams of sulfanilamide over a period of several weeks. The possibility of its occurrence requires that sulfonamides be given with great caution if liver disease be present.

[79] Lyons, R. H., & Balberor, H. Jour. Amer. Med. Assoc. 118: 955. 1942.
[80] Nelson, J. Jour. Amer. Med. Assoc. 119: 560. 1942.

THERAPEUTIC CONSIDERATIONS

The following therapeutic rules emerge from any discussion of the toxic action of sulfonamides:

1. No sulfonamide should ever be employed therapeutically unless its use is clearly indicated. Sulfonamides are not drugs suitable for trivial complaints and should not be employed for a disease in which the diagnosis is in doubt unless the patient's life appears otherwise to be endangered

2. When administering sulfonamides, the physician should be constantly on the alert for signs and symptoms of toxic effects, especially with reference to changes in the blood and in the urine, to avoid the serious complications of granulocytopenia and severe renal damage owing to the precipitation of free or acetylated sulfonamide.

3. Doses should be adequate, but never excessive or too small The most satisfactory guide to adequate dosage is the level of free sulfonamide in the blood.

4 Sulfonamides should be employed for the minimal time necessary. It is extremely unlikely that the continuation of adequate therapy longer than two to three weeks will be of benefit to the patient. It is a wise rule to redouble the search for signs or symptoms of sulfonamide poisoning whenever therapy is continued longer than 10 days

ANTAGONISTS (EXCLUDING *P*-AMINOBENZOIC ACID), DYNAMISTS AND SYNERGISTS OF THE SULFONAMIDES

By Henry Irving Kohn

From the
Department of Physiology and Pharmacology,
Duke University School of Medicine,
Durham, North Carolina

INTRODUCTION

A knowledge of the antagonists and dynamists of the sulfonamides is of importance from both the practical and the theoretical points of view On the one hand, it can suggest methods for increasing the potency of the sulfonamides, and it can also tell us why these drugs are ineffective in certain locations On the other, it is stimulating the investigation of metabolic systems, previously unknown, which play a major role in the anabolic processes of growth and multiplication In this review, antagonists, dynamists, and to some extent, synergists will be discussed under three headings. (1) determination of activity, (2) summary of the literature, and (3) mode of action

DETERMINATION OF ACTIVITY

A survey of the literature shows no agreement upon what constitutes an adequate technic for the quantitative measurement of antagonism or dynamism, most workers being content with a qualitative statement of their results Because of this, it is often difficult to estimate how active a given agent is, or to compare the activity of different agents, or the work of different laboratories. In a general way, the outline for a quantitative method is implicit in the definition of the agents under discussion.[1]

Antagonists increase the concentration of sulfonamide required to produce a chosen degree of inhibition. Therefore, let the fold-increase in the sulfonamide concentration be the measure of antagonism For example, in the absence of *X*, 1 mg. per cent sulfonamide inhibits the rate of growth by 50 per cent. In the presence of *X*, 5 mg. per cent sulfonamide inhibits by 50 per cent. Here the antagonism is 5-fold.

[1] Kohn, H. I., & Harris, J. S. Jour. Pharmacol. 77: 1. 1943.

Dynamists[1] (or potentiators) are not inhibitory *per se*, but do decrease the sulfonamide concentration required to produce a chosen degree of inhibition. Suppose that the addition of *Y* to the medium, although without effect by itself, decreases the required sulfonamide concentration for 50 per cent inhibition from 1 mg. per cent to 0.1 mg. per cent. Here the dynamism would be 0.1-fold. It will be noted that on this scale of measurement, fold-changes greater than 1 indicate antagonism, while those less than 1 indicate dynamism.

Synergists are inhibitory *per se:* their presence decreases the required concentration of sulfonamide. The measurement of synergism is difficult because both the agent and the sulfonamide inhibit growth. A possible method involves the testing of synergists at only certain "functional" concentrations, e.g., those producing inhibitions of 5, 10 and 20 per cent At each of these, the concentration of sulfonamide necessary to produce the chosen degree of inhibition would be determined, and the fold-change calculated

In studying the mode of action in PAB, a number of workers[3-6]; have calculated the SA/PAB ratio, i e , the number of molecules of sulfonamide antagonized by one molecule of PAB, and have found it to be constant over a wide range of SA concentration Although its significance is somewhat different, the SA/PAB ratio can be translated into the terms of fold-antagonism. For example, suppose that the test organism is inhibited 50 per cent in the presence of 10^{-5}M SA and 10^{8}M PAB, the value of the SA/PAB ratio being 1,000. If the PAB concentration is increased 10, or 100, or 1,000 times, the SA concentration must be increased likewise to maintain the chosen degree of inhibition (50 per cent), and the fold-antagonism will increase from 10 to 100, and to 1,000. Note that PAB is effective throughout the entire SA concentration range, and consequently the fold-antagonism can be made almost as great or as small as desired by adjusting the concentration of PAB.[7] Obviously, in this case the fold-antagonism is a less elegant measure of antagonism than the SA/PAB ratio.

When dealing with agents other than PAB, however, the situation is different, and the fold-change provides a suitable measure. The reason for this lies in the fact that the agents known today are *not* effective

---

[1] I am indebted to Prof E. V. Way of the Greek Department for suggesting this word. It will be noted that *dynamist*, *dynamize*, and *dynamism* are analogous to *synergist*, *synergize*, and *synergism*, whereas the awkward *potentiator* is not

[3] Woods, D. D. Brit. Jour. Exp. Path. 21: 74. 1940.

[4] Wyss, O. Proc. Soc. Exp. Biol. Med. 48: 122. 1941.

[5] Rose, H. M., & Fox, C. L., Jr. Science 95: 412. 1942.

[6] Wood, W. B., Jr. Jour. Exp. Med. 75: 369. 1942.

[7] This statement is subject to the qualification that molar concentrations of PAB more than about [illegible] are per se increasingly inhibitory and at the same time gradually lose their antagonistic action.

throughout the entire SA concentration range. In other words, it is always possible to increase the SA concentration to such an extent that the antagonist ceases to be effective no matter how much of it is employed. For this reason, there is an optimal concentration or concentration range for each agent, and, when this is employed in the test, the fold-change calculated is a convenient and highly significant datum.

The design of an experiment to measure the fold-change involves a number of considerations. The general plan of the experiment is to determine the concentration of SA producing a chosen percentage of inhibition, both in the presence and absence of the substance under test. It is to be emphasized that the inhibition is in the *rate of growth,* and should be independent of the number of cells present. The latter condition, perhaps, can not always be achieved in practice (e.g., in the case of organisms excreting PAB into the medium), though it probably can always be approximated. The necessity for thinking in terms of rates instead of cell numbers is illustrated by the following example: if from inocula of 1 cell, two cultures produce respectively 1,000 cells (control) and 500 cells (drug), the average inhibition in the rate of growth for the period is 10 per cent, not 50 per cent, because 10 generations occurred in the control compared to 9 in the test culture.

The following points are suggested for consideration in the design of experiments:

1. The mean rate of growth for some extended period should be measured. Under standardized conditions, this is accomplished most easily by noting the time required for a given inoculum to produce a density of population that can be measured conveniently. The density chosen can be measured by the eye ("just visible turbidity"), by the photometer, by plate counts, etc. Photometric determinations are perhaps the most accurate, and also are convenient since the cultures can be grown directly in the photometer tubes. In any case, the percentage inhibition in the rate of growth is

$$100\left[1 - \frac{\text{hours for control growth}}{\text{hours for experimental growth}}\right].$$

2. The period of growth should involve 12 to 20 generations. Such an extended period has two advantages: it minimizes the effect of the latent period (2 to 5 generations) during which the sulfonamide action is gradually developing; it also provides an adequate opportunity for expression in the case of those agents that act after the latent period is completed.

3. It must be established that control growth is exponential, i.e., the

rate is constant during the period of growth. The necessity for this precaution is illustrated by the data in FIGURE 1, where the number of generations (or divisions) is plotted against the time in hours for three different cultures identified as C (control), No. 2 (2 mg. per cent SA), and No. 5 (5 mg. per cent SA). Although hypothetical, the data are representative of what can be obtained experimentally. For material illustrative of this and other points discussed here see Kohn and Harris.[8] At zero time the cultures each contained 100 organisms per cc. The percentage inhibition of the rate of growth calculated at various times is shown in TABLE 1 The table shows that the calculated inhibition will be too small if the period of growth is too short. The table also shows the fallacy of arbitrarily selecting a certain time at which to compare the amounts of growth. For example, when the comparisons are made at 35 hours after inoculation, the fallacious conclusion is drawn that 2 mg. per cent sulfonamide does not inhibit growth.

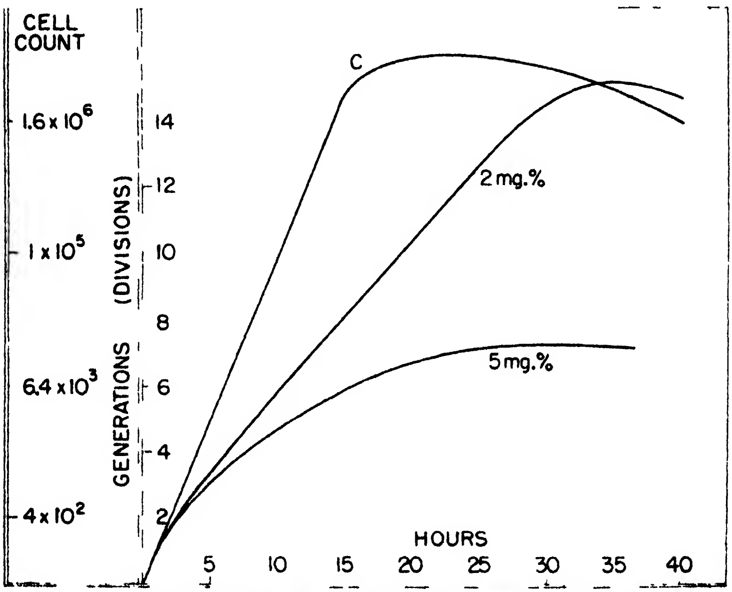


FIGURE 1 Growth as a function of time On the ordinate, both cell counts and generations (divisions) are given, the inoculum at zero time being 100 cells Curve C is for the controls, the others for cultures containing 2 and 5 mg per cent SA For comment, see point 3 in the text

[8] Kohn, H. I., & Harris, J. S. Jour. Pharmacol. 73: 343. 1941.

TABLE 1

RATE OF GROWTH AND PERCENTAGE INHIBITION CALCULATED FOR THE DATA OF FIGURE 1

Culture	Number of generations	Growth period	Rate of growth	Mean inhibition
		hours	*generations/hour*	%
Control	1	1	1	—
No 2	1	1	1	0
No 5	1	1	1	0
Control	6	6	1	—
No 2	6	10 5	0.57	43
No 5	6	15.5	0 38	62
Control	14	14	1	—
No 2	14	28	0 50	50
No 5	14	?	0	100
Control	(15)	35	(0 43)	—
No 2	15	35	0 43	(0)

4 The degree of inhibition at which the comparisons are to be made, called the *endpoint,* can be selected with assurance only after a preliminary trial, particularly when peptone is present The data shown in FIGURE 2 illustrate this situation, where the rate of growth as a percentage of the control (no sulfonamide) is plotted as a function of the sulfonamide concentration (log scale) The right-hand pair of curves was made with sulfanilamide, the open circles representing a synthetic medium, the solid circles the same medium + 1 per cent peptone The two curves are almost parallel, and consequently it makes some but not much difference where the endpoint is chosen Using an endpoint of 75 per cent inhibition, the antagonism of peptone (in this particular case) is 3-fold, at 50 per cent 1 5-fold On the other hand, in the case of the left-hand pair of curves made with sulfathiazole, the selection of the endpoint is of paramount importance At 50 per cent the antagonism is 1 8-fold, while at 77 per cent it is about 30-fold[6] This example shows that the selection of the endpoint may depend upon the sulfonamide drug under test, and also upon the antagonist Thus far only peptone is known to possess this peculiarity, and a 50 per cent endpoint is satisfactory in its absence

5 The effect of the test substance upon the rate of growth must be determined in the absence of sulfonamide If the rate is unaffected, as is the case with PAB but very few others, a simple endpoint such as "no

[6] The fold-antagonism with peptone is usually larger than this when peptone is added to a synthetic medium In the above example, the basal medium contained substances that did not permit the maximal activity of the peptone to be exhibited

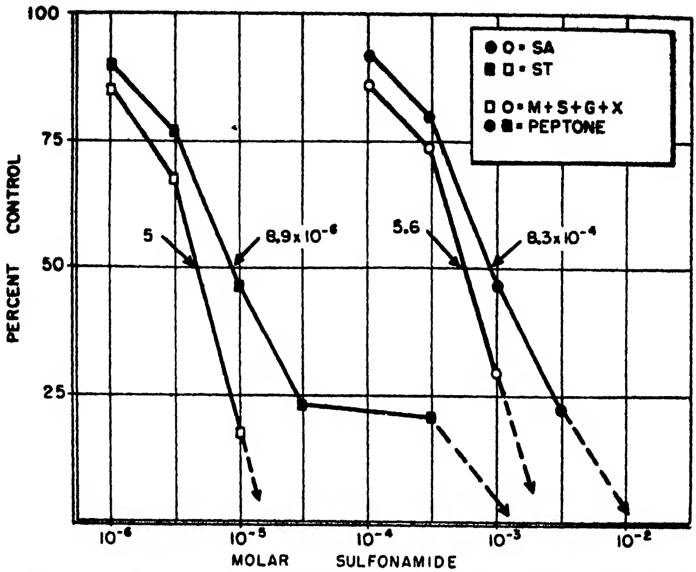


FIGURE 2. Rate of growth (as percentage of the control) plotted as a function of the sulfonamide concentration (log scale). The right-hand pair of curves was made with sulfanilamide, the left with sulfathiazole. In each pair the open circles were obtained in the basal medium (salt, glucose, methionine, serine, glycine, and xanthine), the solid circles in basal medium × 1 per cent peptone. For comment see points 4 and 6 in the text.

visible turbidity in two days followed by turbidity on the third" will suffice.[5, 6] If, however, the substance under test changes the control rate of growth, the experiment becomes more complicated. For example, let the endpoint be 50 per cent. In the *absence* of test substance, supposing the controls to grow out in 10 hours, the SA concentration (SA-c) producing 20-hour growth is determined. In the *presence* of test substance, the controls grow out in 5 hours, hence the concentration of SA (Sa-t) producing 10-hour growth is determined. The fold-change will be: (Sa-t)/(SA-c). Examples can be found in several articles.[1, 10, 11] Such considerations are of the utmost importance in dealing with peptone, for example, which under the standardized conditions employed by Harris and myself can reduce the time for 22 generations of *Escherichia coli* from about 18.5 to 8 hours.

6. The optimal concentration range for the test substance should

[10] Harris, J. S., & Kohn, H. I. Jour. Pharmacol. 73: 383. 1941
[11] Wyss, O., Grubaugh, K. K., & Schmelkes, F. C. Proc. Soc. Exp. Biol. Med. 49: 618. 1942. Wyss, O., Strandskov, F. B., & Schmelkes, F. C. Science 96: 236. 1942.

be determined, and the fold-change calculated with respect to it. Results might be briefly summarized as follows:

(adenine: 10^{-4}) 0.25-F (sulfanilamide: 10^{-4}).

This signifies that 10^{-4}M adenine potentiates sulfanilamide 0.25-fold, i.e., that equal degrees of inhibition are produced by 10^{-4}M sulfanilamide or by 10^{-4}M adenine + 0.25×10^{-4}M sulfanilamide

7. The temperature should be 37–38°, though the use of others is also desirable, since activity may depend upon temperature.[12-14]

8. The pH of the medium should be 7–8, and accurately known.[15, 16] The medium should be synthetic if possible, and its composition given in detail.

9. When possible, a readily accessible test organism should be employed, as from the American Type Culture Collection.

10. When both metabolic (e.g., oxygen consumption) and growth-rate experiments are done on the same organism, either the concentration of the drug or of the test substance should be held constant throughout, while the other is varied.

11. In conclusion, it may be pointed out that the interpretation of such experiments is complicated by the fact that the organisms may change during the course of the experiment For example, they may develop some resistance to the SA present, or change metabolically in other ways. Such changes no doubt are implied when, during the course of an experiment, the growth of organisms is slowed, then stopped, but, after some time, begins once more.

If the action of an agent as a function of time is to be studied, other methods than the above must be employed. The usual methods involve either periodic sampling for viable counts, or following the increase in turbidity with the photometer. The former method is laborious. The latter is likely to be limited to a restricted period of growth. What seems to be a better method than either is the K_{sd} technic[1, 8] illustrated by the data in FIGURE 3. A turbidity is selected that is readily measured in the photoelectric colorimeter and at which growth is still exponential. At zero time, the photometer tubes are inoculated with serial dilutions such that the first will reach the endpoint of turbidity in 1 generation, the second in 2, the third in 3, etc. The time for each culture to reach the turbidity endpoint is plotted as in FIGURE 3

[12] White, H. J. Jour. Bact. 38: 549. 1939.
[13] Weld, J. T., & Mitchell, L. C. Jour. Bact. 38: 335. 1939.
[14] Kalmanson, G. M. Jour. Bact. 40: 817. 1940.
[15] Bell, P. H., & Roblin, R. O., Jr. Jour. Amer. Chem. Soc. 64: 2905. 1942.
[16] Schmelkes, F. C., Wyss, O., Marks, H. C., Ludwig, B. J., & Strandskov, F. B. Proc. Soc. Exp. Biol. Med. 50: 145. 1942.

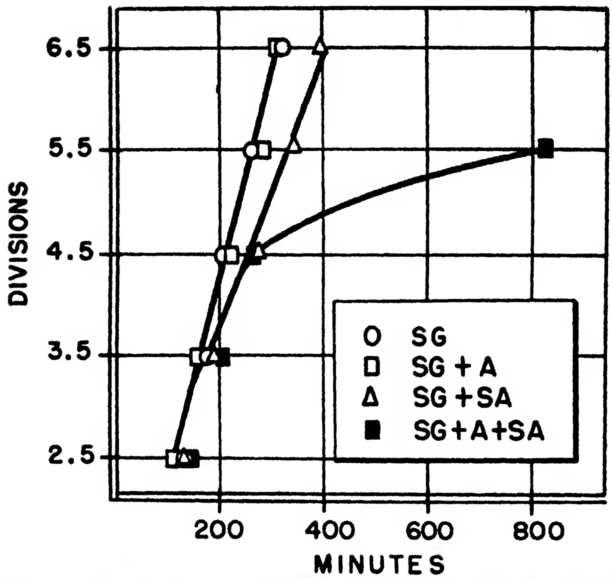


FIGURE 3 Number of generations (divisions) as a function of time The data were determined by the K_{al} technic (see text). The four cultures compared are basal medium SG plus adenine SG + A, + sulfanilamide SG + SA, and plus adenine and sulfanilamide, SG + A + SA. Note the constant rate of growth of the controls, the latent period before inhibition occurs, and the subsequent adenine dynamism (from Kohn and Harris, ref 1)

It is of interest that the general approach outlined above for the measurement of antagonism and dynamism can be applied to the measurement of resistance to drug action, or fastness.[17] In this case, the resistance of the organism is measured in terms of the SA concentration required to produce a 50 per cent inhibition in the rate of growth before and after the training period, and the fold-change calculated.

SURVEY OF THE LITERATURE

This section deals with a summary of the established facts concerning antagonists and dynamists without reference to theory or interpretation When abstracting papers, the original data were evaluated from the point of view presented in the preceding section, and on occasion the author's conclusions have not been accepted. Although this review, deals primarily with substances other than PAB, it has been found necessary to include some mention of this compound.

[17] Harris, J. S., & Kohn, H. I. Jour. Immunol. 46: 189. 1943. In press

The substances under review fall into two categories: biological extracts of more or less unknown composition, and pure chemical compounds. The extracts will be discussed according to source, the compounds according to chemical structure. The synergists are listed briefly in a separate section, as also are certain animal experiments.

Biological Extracts

BACTERIA.—The work of Green[18] and of Stamp[19] culminated in the discovery by Woods[9] of PAB, which antagonizes all sulfonamides in all species and media. Subsequently, Green[20] conceded that the antagonistic action of his "P" factor was due to contamination with PAB. Rubbo and Gillespie[21] showed PAB, but not the ortho or meta compounds,[22] to be a growth factor for *Cl. acetobutylicum*. This has been extended to species of *Lactobacillus* and *Acetobacter*. *P*-aminophenyl acetic acid is a growth factor but not an antagonist[22, 23]; *p*-aminobenzamide is a weak inhibitor like sulfonamide.[24] MacLeod[25] found some species (*Staphylococcus aureus* and *Pneumococcus*, but not *E coli* and Group D *Streptococcus*) to release antagonists into the medium, resistant pneumococci producing much more than the parent strain. This antagonist was largely if not entirely PAB, according to recent microbiological assays,[26] which also showed that the development of resistance in certain other species (e g, *E. coli*) is not accompanied by enhanced production of PAB. Mirick[27] found a soil bacillus producing an antagonist, which, though unidentified, is not PAB

YEAST.—Woods[9] isolated his most potent material from yeast The actual isolation of *p*-aminobenzoic acid was achieved by Rubbo and Gillespie,[21] and especially by Blanchard,[28] who indicated a concentration of as much as 0 5 mg per cent. Yeast extracts contain other antagonists that, for example, do not have free amino groups.[29]

PEPTONE.—This was discovered to be an antagonist by Lockwood[30] and by Fuller and coworkers[31] Weld and Mitchell,[32] using rabbit instead of human serum (neopeptone broth, hemolytic *Streptococcus*), showed that the action varied with the temperature: 37 5° C., antag-

[18] **Green, H. N.** Brit Jour. Exp Path. **21**: 38. 1940.
[19] **Stamp, T. C.** Lancet **2**: 10 1939
[20] **Green, H. N., & Bielshowsky, F.** Brit. Jour Exp Path **23**: 1. 1942
[21] **Rubbo, S. D., & Gillespie, J. M.** Nature **146**: 838 1940
[22] **Rubbo, S. D., & Gillespie, J M.** Lancet **242**: 36. 1942
[23] **Landy, M.** Proc Soc. Biol Med **46**: 59 1941
[24] **Hirsch, J.** Science **96**: 139 1942.
[25] **MacLeod, C. M.** Jour Exp. Med **72**. 217. 1940
[26] **Landy, M., Larkum, N. W., Oswald, E. J., & Streightoff, F** Science **97**: 265. 1943.
[27] **Mirick, G. S.** Jour Bact **45**: 66. 1943
[28] **Blanchard, K. C.** Jour. Biol. Chem. **140**: 919. 1941.
[29] **Loomis, T. A., Hubbard, S., & Neter, E.** Proc. Soc. Exp. Biol. Med. **47**: 159. 1941.
[30] **Lockwood, J. S.** Jour. Immunol. **35**: 155. 1938.
[31] **Fuller, A. T., Colebrook, L., & Maxted, W. R.** Jour Path Bact **51**. 105 1940

onism; 39° C., potentiation. The growth-promoting action of peptone does not account for its antagonism in *Streptococcus*[32] or in *E. coli.*[8] Its content of PAB is too small to account for its antagonistic action.[1, 33] Using a strain of *E. coli* that grows well on inorganic salts and glucose, detailed quantitative analysis showed peptone to contain a number of antagonists and at least one dynamist.[1, 8, 10] The antagonists may be divided into two groups:

1. *Equally antagonistic to all drugs.* Using 1 per cent peptone, the antagonism is 8- to 15-fold measured at the 50 per cent endpoint. Methionine, serine, glycine, allothreonine (but no other known naturally occurring amino acid), xanthine and guanine are the compounds that probably constitute this group. Traces of PAB present in peptone count in this group. Excluding methionine, the members of this group were called P-1.

2. *Antagonistic to SP, ST and SD much more than SA; antagonism evident only when the rate of growth is inhibited by more than 65 per cent* Measured at the 75 per cent endpoint, the antagonism is about 3-fold against SA, and 30-fold against SP, ST, and SD. The best source of P-2 is pancreas It is not a known naturally occurring amino acid, a heat labile protein, nor insulin. Its action is additive with that of group 1. A fairly potent preparation can be obtained by extracting defatted pancreas with hot 95 per cent alcohol, evaporating the extract to dryness, and dissolving the residue in water (Harris and Kohn, unpublished data)

White and coworkers[34] determined the relative activity of several sulfonamide drugs for three strains of *E coli* in the presence and absence of several amino acid hydrolysates and peptone. The endpoints chosen were different and not precisely specified for the different media, but, in any one medium, were the same for all the drugs tested. The interpretation of their data is complicated by the fact that one cannot tell whether their endpoints fell on or off the plateau (see FIGURE 2 and page 507, comment 4). They found that the relative potencies of ST, SD, and SP were constant and roughly equal in the basal medium, or when peptone, protein hydrolysate, methionine or PAB were added. The relative potency of SA compared to the heterocyclic derivatives, however, was low in the basal medium, but tended to approximate equality in the presence of tryptose peptone or PAB. They suggested that "the anti-drug effect of either PAB or some factor in peptone is

[32] Lynch, H. M., & Lockwood, J. S. Jour. Immunol. 42: 435. 1941.
[33] Hill, J. H., & Mann, E. F. Jour. Urol. 47: 522. 1942.
[34] White, H. J., Litchfield, J. T., & Marshall, E. K., Jr. Jour. Pharmacol 73: 104. 1941

exerted to a much less extent on this drug (SA) than on its heterocyclic derivatives." This is an independent formulation of the antagonists labeled as group P-2, above. The technical explanation for their results can be seen by inspecting FIGURE 2 and comparing the right-hand members of each pair of curves. As Kohn and Harris (table 1 in ref. 8) have pointed out, in peptone media the presence of the plateau tends to equalize the potency of SA and its heterocyclic derivatives when the endpoint is around 75 per cent, but does not when the endpoint is, say, 30 per cent. They found no plateau upon adding PAB, however, which may indicate a difference between the strains of *E. coli* used in the two laboratories.

URINE, PUS, AND TISSUES.—Water-soluble antagonists can be obtained from all of these. MacLeod[25] found the activity of extracts to be increased by mild acid hydrolysis. When the medium was made up to contain methionine, serine, glycine and xanthine, however, of rabbit tissues, only the pancreas showed appreciable antagonism with *E coli* [1] A factor in mouse urine and another in erythrocytes have been partially purified by Fuller and coworkers [31]

PROTEINS—The plasma proteins adsorb the sulfonamides according to Schonholzer[36] and Davis and Wood.[37] For detailed studies on the relation between structure and adsorption, see Shannon[38] and Fisher and coworkers,[39] who found the following percentages of sulfonamide bound by dog albumin (4 per cent) solutions: SG, 6 per cent; PAB, 7 per cent; SA, 10 per cent; SD, 17 per cent; SP, 30 per cent; ST, 60 per cent.

Chemical Compounds

Amino Acids

METHIONINE.—The antagonistic action of methionine was discovered independently in two laboratories.[10, 35, 40] Using *E coli* grown in a salt glucose medium, both groups showed that maximal activity occurs in the range from 10^{-5} to 3×10^{-4}M, and that neither choline, cystine, homocystine nor any other amino acid is active [41] Bliss and Long found high methionine concentrations to be inhibitory. Harris and Kohn reported that the antagonism is 3- to 6-fold against SA, SP, ST, and SD; that the natural isomer is ten times more active than the unnatural; that neither is oxidized, deaminated, nor decarboxylated;

---

[35] Kohn, H. I., & Harris, J. S. Amer Jour Physiol 133: 554 1941.
[36] Schonholzer, G. Klin. Wchns
[37] Davis, B. D., & Wood, W. B. Exp Biol Med. 51: 283. 1942.
[38] Shannon, J. A. Ann. N. Y Acad Sci. 44: 455. 1943.
[39] Fisher, S. H., Troast, L., & Shannon, J. A. Jour. Pharmacol. 1943. In press.
[40] Bliss, E. A., & Long, P. H. Bull. Johns Hopkins Hospital 69: 14. 1941.
[41] For this section, conversion between molarity and mg. per cent can be approximated by assuming a molecular weight of 150. The 10^{-5}M equals 15 mg. per cent and 10^{-4}M equals 1 5 mg per cent

that methionine, but not PAB, antagonizes the inhibitions caused by ethionine and norleucine, with which it apparently competes; and that methionine added to peptone medium is without effect.

Methionine became an essential growth factor for *E. coli* when the strain was subcultured in the presence of both methionine and SA.[42]

Strauss and coworkers[43] confirmed the methionine effect in *E. coli* They also reported that resistant strains of *Staphylococcus aureus* could use methionine as a weak antagonist, although the parent strains could not The authors did not comment upon the fact that the casein hydrolysate present in the medium should have furnished the equivalent of 2×10^{-3}M methionine, whereas they obtained their effects by the addition of 1.7×10^{-4}M. Snell and Mitchell[44] reported that methionine is not an antagonist for lactobacilli, but likewise added methionine to a medium already containing much of it

Wyss and coworkers[45] synthesized a number of PAB derivatives with inhibitory activity, of which the most active was 2-Cl-PAB. It was found that 2-Cl-PAB was more sensitive to antagonism by methionine than are the sulfonamides, and that it was quite inactive in complex media

GLYCINE, DL-SERINE AND DL-ALLOTHREONINE —These substances enhance the antagonistic action of methionine by 1.6-fold against SA and ST in *E. coli* grown in salt-glucose medium [1, 46] They are without effect in the absence of methionine or in the presence of peptone In combination with methionine, each at 4×10^{-5}M, they are as active as a mixture of all known, naturally occurring amino acids

GLUTAMIC ACID —At 10^{-7}M, this acid was reported[33] to antagonize SA in a strain of *E. coli* grown on salt-glucose-asparagine medium. When due allowance for growth stimulation is made, the effect probably will be small

Purines

The action of the purines varies with the species and the experimental conditions, so that even in the same organism a given purine can be both antagonist and dynamist. The optimal concentration range of the four active compounds, xanthine, guanine, hypoxanthine and adenine, is 10^{-5} to 10^{-4}M.

It is of some interest that purine-PAB relationships can be demonstrated in the absence of SA. Harris and Kohn[47] showed that when a

[42] Kohn, H. I., & Harris, J. S. Jour. Bact. 44: 717. 1942.
[43] Strauss, E., Dingle, J. H., & Finland, M. Jour Immunol 42: 331. 1941.
[44] Snell, E. E., & Mitchell, H. K. Arch. Biochem. 1: 93 1942.
[45] Wyss, O., Rubin, M., & Strandskov, F. B. Proc. Soc Biol Med 52: 155. 1943
[46] Kohn, H. I., & Harris, J. S. Federation Proc. 1: 46 1942.
[47] Harris, J. S., & Kohn, H. I. Jour Biol Chem 141: 989. 1941.

strain of *E. coli* became resistant to sulfanilamide, it became sensitive to hypoxanthine so that 10^{-4}M inhibited its rate of growth by 50 per cent. This inhibition was completely antagonized by 10^{-7}M PAB or by 10^{-4}M methionine. More recently Landy and Streightoff[48] have shown in the case of *Acetobacter suboxydans,* which requires PAB as a growth factor, that the addition of purine increases growth at low concentrations of PAB, but not at high.

The complex relations between sulfonamide and purine were first shown in *E. coli,* using a simple medium. It was found[1,47] that dynamism occurs in salt-glucose medium: 0.6-fold for xanthine and guanine, 0.25-fold for adenine and hypoxanthine. But in salt-glucose-methionine medium, xanthine and guanine show 1 8-fold antagonism against SA, SP, ST, and SD; adenine and hypoxanthine show 0.33-fold dynamism. The addition of peptone to the medium cancels these effects. In *Pseudomonas fluorescens,* adenine and xanthine antagonize SA; in *Mycobacterium* sp., they dynamize when methionine is present.

Using lactobacilli grown on hydrolyzed casein plus supplements at 30° C.,[44] the four purines were found to be antagonistic. Suboptimal amounts of PAB or of an unknown factor had to be present in order to show the antagonism in *L. arabinosus* and *pentosus.*

Martin and Fisher[49] reported that adenine sulfate (0 8 mg per gram) antagonized the chemotherapeutic action of SA (2 mg. per gram) or of SD, ST, and SP (4 mg. per gram) in the mouse infected with hemolytic streptococci. Guanine was without effect, though *in vitro* both guanine and adenine are growth factors. Adenine sulfate at 1 mg. per gram was not lethal to untreated or infected mice, but it should not be concluded that this result establishes adenine as a nontoxic substance in these experiments.

Slight antagonistic action against sulfaguanidine, but not against other sulfa drugs, was reported by Strauss and coworkers[45] for 10 mg. per cent sodium nucleate, or a mixture of 1 mg. per cent uracil and adenylic acid and 0.25 per cent pyruvic acid. *Staphylococcus aureus* was employed, grown in a casein hydrolysate medium

Nicotinic Acid and Coenzymes

Antagonism by cozymase in *Staphylococcus aureus* was first reported by West and Coburn,[50] who grew their own strain in Knight's medium (gelatine hydrolysate plus supplements) without thiamin They found nicotinic acid to be inactive at the single concentration

Landy, M., & Streightoff, F. Proc. Soc. Exp. Biol. Med. 52: 127. 1943.
Martin, G. J., & Fisher, C. V. Jour. Biol. Chem. 144: 289. 1942
West, R., & Coburn, A. F. Jour. Exp. Med. 72: 91. 1940.

tested, which was not optimal for growth. Cozymase was obtained as a crude extract of erythrocytes or from yeast (purity 25 per cent ?), and lost its activity following 30 minutes at 120° C., pH 10. The antagonisms reported appear greater than could be accounted for by stimulation of growth, but no data on the action of cozymase alone were presented. They suggested that sulfapyridine inhibits the synthesis of cozymase.

Spink and coworkers[51] using Gladstone's synthetic medium, concluded that cozymase does antagonize SA and SP, but not ST or SD.; but the experiments seem inconclusive. Strauss and coworkers,[43] using a casein hydrolysate medium which contained 1 μgm./ml. of nicotinamide, found that addition of cozymase (0.1 to 15 μgm./ml.) or of nicotinamide (up to 200 μgm. /ml.) to be without effect.

Wood and Austrian,[52] using Gladstone's synthetic medium, found 10^{-3}M nicotinamide and 10^{-6}M cozymase to be equivalent in growth promotion and antagonism. The effect is nonspecific since cozymase antagonized thionine equally well. The antagonism seemed clear cut, but cannot be accurately assessed because simultaneous controls for growth stimulation (control plus antagonist but without sulfonamide) were not reported.

In the case of *Lactobacillus arabinosus*[53] grown on hydrolyzed casein, which requires nicotinamide as a growth factor, 0.2 mg. per cent SP was antagonized by 0.1 mg per cent nicotinamide, by 0.5 mg. per cent cozymase, or by 0.5 mg. per cent nicotinamide nucleoside. In these determinations, no allowance was made for stimulation in the absence of SP.

In the case of the dysentery bacillus, there is evidence from both respiration and growth experiments that nicotinamide, which is a specific growth factor, antagonizes ST and SP.[54, 55] When tested in a synthetic medium, raising the nicotinamide from 0.0006 mg. per cent to 0.0024 mg. per cent antagonized 0.2 mg. per cent ST; but raising it to 0.01 mg. per cent did not antagonize 1 mg. per cent ST. In respiration experiments with cells suspended in phosphate buffer plus glucose containing 0.1 to 0.3 mg. per cent nicotinamide, 85 per cent inhibition was obtained with 10–30 mg. per cent SP or ST. The acetylated form of SP was also active, thus indicating a difference in mechanism from that generally operating in the inhibition of growth. It is possible

[51] Spink, W. W., Vivino, J. J., & Mickelsen, O. Proc. Soc. Exp. Biol. Med. 50: 31. 1942.
[52] Wood, W. B., & Austrian, R. Jour. Exp. Med. 75: 383. 1942.
[53] Tepley, L. J., Axelrod, A. E., & Elvehjem, C. A. Jour. Pharmacol. 77: 207. 1943.
[54] Dorfman, A., Rice, L., Koser, S. A., & Saunders, F. Proc. Soc. Exp. Biol. Med. 45: 750 1940.
[55] Dorfman, A., & Koser, S. A. Jour. Inf. Dis. 71: 241. 1942.

that lower sulfonamide concentrations would be effective if the nicotinamide concentration were decreased. A complication in these experiments is the fact that inhibition of respiration with 10–30 mg. per cent sulfonamide is possible only when nicotinamide-starved cells are incubated with sulfonamide for one hour before respiration is initiated by the addition of glucose and nicotinamide. It is therefore quite possible under the usual conditions of adequate nutrition that the sulfonamides inhibit growth but not respiration. In any event, the effect of the sulfonamides upon the respiration of the dysentery organism seems to be a special case, at least for the present. It must be pointed out, however, that analagous nicotinamide-starvation experiments have not been done on other species

Urea and Asparagine

Tsuchiya and coworkers[56-59] reported that urea (1 7 per cent) dynamizes the action of SA, ST and SD in *E. coli* grown in salt-glucose medium containing methionine or PAB. It also dynamizes the action of ST on resistant strains of *Staphyllococcus aureus* The fold-change cannot be calculated from their data. The authors feel that the mode of action is not settled and "would like to stress that we do not know that urea acts against inhibitors" Fox[60] found that such high concentrations of urea inhibited the growth of his strain of *E. coli*, but had no dynamistic action. Schmelkes and Wyss[61] using another strain of *E coli* determined the dynamism as 0.3-fold for 1 per cent urea, and 0 2-fold for 2 per cent asparagine.

Synergists

Pyridium[62] was active in infusion broth against *E. coli;* azochloramid[63, 64] against the *Streptococcus, Pneumococcus,* and *E. coli* in simple and complex media, and in the presence of PAB or resistant organisms[65] The conclusion that it acts specifically against sulfonamide antagonists is unwarranted. Ethionine and norleucine act against *E. coli* in salt-glucose medium, but not in the presence of methionine or peptone[10]

---

[56] Tsuchiya, H. M., Tenenberg, D. J , Clark, W. G., & Strakosch, E A Proc Soc Exp Biol Med. 50: 262. 1942.
[57] *Ibid.* 51: 245. 1942.
[58] *Ibid.* 51: 247. 1942.
[59] Tsuchiya, H. M. Personal communication.
[60] Fox, C. L. Personal communication. 1943.
[61] Schmelkes, F. C., & Wyss, O. Personal communication 1943.
[62] Neter, E., & Loomis, T. A. Urol. Cutan. Rev. 45: 2. 1941.
[63] Neter, E. Proc. Soc Exp. Biol. Med. 47: 303. 1941.
[64] Neter, E. Jour. Pharmacol. 74: 52. 1942.
[65] Schmelkes, F. C., & Wyss, O. Proc. Soc. Exp. Biol. Med. 49: 263 1942

Animal Experiments

In concluding this section, it is of interest to recall that the sulfonamides, at approximately clinical levels, can inhibit the growth of the rat. The principal factors responsible seem to be: inhibition of the synthesis of vitamins by the intestinal flora[66]; specific histotoxic effects[67, 68]; and a specific inhibition of appetite.[69] Of particular interest in the present discussion is the discovery that all of these actions can be antagonized by the addition of various supplements to the basal purified rations employed in the experiments. Such antagonists include meat, meat products, liver extract, PAB, folic acid, biotin, peptone, yeast extract, and feces.[68, 70-73]

MODE OF ACTION

Since a sulfonamide, like other drugs, may be expected to act at an increasing number of qualitatively different loci as its concentration is raised, the mechanisms under consideration will be limited to those affected at or below safe clinical levels, the upper limit of which will be taken as 12 mg. per cent or 5×10^{-4}M. This qualification, of course, does not exclude experiments in which the upper limit has been crossed on purpose, as in demonstrating the activity of an antagonist.

A survey of the data presented in the preceding section shows that only a beginning has been made in the investigation of antagonists and dynamists, and consequently we may expect little in the way of specific theory to account for their actions. Furthermore, crucial experiments whose biochemical interpretation is unambiguous, such as can be done with isotopes, have not yet been performed. The discussion which follows, therefore, is of necessity couched in rather general terms.

Agents affecting the action of the sulfonamides may do so in a number of ways. They may influence the penetration of the drug into the bacterium, or the interaction of the drug at the sensitive locus. The effects of pH and of variations in structure of the sulfonamides themselves no doubt can be traced to such mechanisms as these; and they are discussed by Roblin and Bell[74] elsewhere in this article. When dealing with the usual antagonists and dynamists, however, it is difficult to decide whether they operate in this fashion, or otherwise as dis-

[66] Black, S., Overman, R. S., Elvehjem, C. A., & Link, K. P. Jour. Biol. Chem. 145: 137. 1942.
[67] Mackenzie, J. B., & Mackenzie, C. G. Federation Proc. 1: 122. 1942.
[68] Daft, F. S., Ashburn, L. L., & Sebrell, W. H. Science 96: 321. 1942.
[69] Harris, J. S., & Kohn, H. I. Jour. Pharmacol. 73: 56. 1943.
[70] Nielsen, E., & Elvehjem, C. A. Jour. Biol. Chem. 145: 713. 1942.
[71] Welch, A. D. Federation Proc. 1: 170. 1942.
[72] Light, R. F., Cracas, L. J., Olcott, C. T., & Frey, C. N. Jour. Nutrition 24: 427. 1942.
[73] Welch, A. D., & Wright, L. D. Jour. Nutrition. 1943. In press.
[74] Roblin, R. O., Jr., & Bell, P. H. Ann. N. Y. Acad. Sci. 44: 449. 1943.

cussed below. Because it seems more profitable to do so, we shall disregard these factors and assume that metabolic mechanisms are at work. The two exceptions to this include urea and asparagine, concerning which practically nothing is known The high concentrations necessary to produce dynamism together with the well known peptizing action of urea, however, suggest that specific metabolic interactions cannot be at work Lastly, the antagonism due to proteins that enter into loose combinations with the sulfonamides in the culture medium or body fluid, though of importance, will require no further comment here

Any theory to account for the mode of action of an antagonist or dynamist must be consistent with certain cardinal facts relating to the action of the sulfonamides in general At present, the following may be mentioned:

1. The drugs act after a latent period during which growth must occur [8, 75, 76]
2. The drugs themselves seem to be the active agents.[31, 77]
3 There is no appreciable change in the drug concentration of the medium during the course of experiments involving the usual numbers of bacteria.[8, 60, 78]
4. Growth is affected primarily; respiration may be affected secondarily[8, 11, 79-83], but also see MacLeod,[84] whose data suggest complications
5. *P*-aminobenzoic acid is a specific and complete antagonist Apparently it is an essential metabolite which most bacteria can synthesize, but which some cannot [3, 85] It is assumed to be essential for the growth of susceptible cells, and that the enzyme reaction in which it is involved is subject to competitive inhibition by the sulfonamides.

In order to account for the antagonism of methionine in *E coli*, Harris and Kohn[10] suggested an extension of the PAB system as illustrated in FIGURE 4. The reaction or reactions in which PAB is involved are called the primary reactions, and they constitute the locus at which the sulfonamides react. The primary reaction gives rise to the primary products that are pictured as entering into secondary reactions These in turn give rise to the secondary products to whose concentration, or rate of production, growth may be proportional. In

---

75 Bliss, E. A., & Long, P. H. Jour. Am. Med. Assoc. 109: 1524. 1937
76 Wolff, L. K., & Julius, H. W. Ann. Inst. Pasteur 62: 616. 1939.
77 Lowell, F. C., Strauss, E., & Finland, M. Jour. Immun. 40: 311. 1941.
78 Wyss, O.; Fox, C. L. Personal communications.
79 Kempner, W., Wise, B., & Schlayer, C. Amer. Jour. Med. Sci. 200: 484. 1940.
80 Grieg, M. E., & Hoogerheide, J. C. Jour. Bact. 41: 557. 1941.
81 Kempner, W. Federation Proc. 1: 46. 1942.
82 Schlayer, C. Federation Proc. 1: 78. 1942.
83 Clifton, C. E., & Loewinger, I. E. Proc. Soc. Exp Biol. Med 52: 225. 1943
84 MacLeod, C. M. Proc. Soc. Exp. Biol Med. 41: 215. 1939.
85 Fildes, P. Lancet 1: 955. 1940.

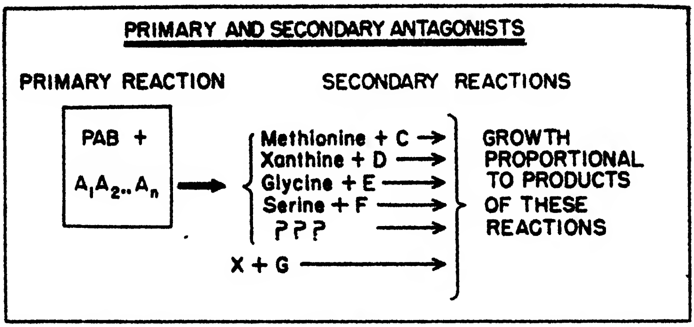


FIGURE 4 Scheme suggested to show the relationships between PAB and the reactions dependent upon it.

addition, one must picture independent reactions (such as X + G in FIGURE 4) which also are essential for growth, but which are independent of the primary, and therefore of PAB. Lastly one may conceive of pre-primary reactions, not indicated in the figure, in which PAB and its fellow reactants are produced.

The diagram states that an inhibition of the primary reaction in turn inhibits growth because the products of the secondary reactions are no longer produced in adequate amounts Furthermore, complete restoration should be possible by the addition of enough PAB. The diagram also predicts that complete restoration should be possible without the addition of PAB by supplying the lacking primary products and thereby re-establishing the secondary reactions, or by supplying the secondary products themselves. The basic plan presented in FIGURE 4 is a simple one, and from it others of much greater complexity can be derived.

To picture the mode of action of methionine, let us assume that it is a primary product When SA inhibits the primary reaction, the synthesis of the primary products will be inhibited, though each not to the same extent. Suppose that at low concentrations of SA only the production of methionine is inhibited. Under these circumstances, the addition of enough methionine to the medium will completely antagonize SA. Higher concentrations of SA will also inhibit the production of other primary products. We may suppose that when the production of methionine is inhibited 100 per cent, that of xanthine is decreased by 50 per cent. Methionine then is no longer a complete antagonist, and xanthine will not act in the absence of methionine. This, in fact, is what is found experimentally.

Methionine is called a secondary antagonist because it takes part in a secondary reaction. Glycine, serine, allothreonine, xanthine and guanine have also been called secondary antagonists, though it is possible that some are actually tertiary ones.

The position of methionine in the scheme is supported by several additional pieces of evidence The structurally related ethionine inhibits growth and is readily antagonized by methionine, but not by PAB. This suggests that the locus of methionine action follows that of PAB, and is not concerned with the synthesis of the latter. It was also found[42] that when strains are subcultured in the presence of both sulfanilamide and methionine, methionine becomes an essential growth factor.

The scheme proposed to explain the results with *E. coli* is consistent with the cardinal facts enumerated above. The latent period of action is attributed to the fact that the cells have stores of essential materials, and until these have been consumed in growth, no inhibition will occur It permits the drugs *per se* to be the active agents, and it does not require that appreciable amounts of them shall be consumed. It directs its attention primarily to the anabolic phases of metabolism, and therefore requires no immediate effect upon respiration Finally, it assigns the key role to PAB

The scheme does not explain why all four purines are dynamists in the absence of methionine, whereas xanthine and guanine are antagonists in its presence Evidently structural specificity is of importance here, substitution at the 2-position being associated with potential antagonism, but at the 6-position with dynamism These relationships are true of *E. coli*, but not necessarily of other organisms

The scheme also fails to account for the differences between the heterocyclic derivatives and sulfanilamide in the presence of peptone The plateaus obtained with the heterocyclic derivatives, as illustrated in FIGURE 2, have been interpreted to mean that the latter can affect reactions that sulfanilamide can not. These reactions become limiting for growth, and hence make themselves known, only when the rate of growth has been inhibited by 75 per cent or more The scheme could be amplified to include these data, but its complexity would be considerably increased.

The scheme accounts for two other experiments in which no sulfonamide was involved; in the first case somewhat vaguely, in the second rather specifically. It will be recalled that hypoxanthine and adenine dynamize sulfanilamide, but have no effect upon growth in its absence. When the strain of *E. coli* was made resistant (growth occurred in 130

mg. per cent sulfanilamide), both these purines became inhibitory *per se*, and their actions were antagonized by either PAB or methionine. These experiments together with those discussed above definitely established metabolic relationships between PAB, methionine and several other amino acids, and the purines in *E. coli.* The second case was recently reported by Landy and Streightoff,[48] employing *Acetobacter suboxydans* which requires PAB as a growth factor. FIGURE 4 predicts in such a case that at suboptimal concentrations of PAB the addition of purine will increase growth, though it will not bring it to the maximum obtainable with optimal PAB. Furthermore, when PAB is optimal, the addition of purine will have no effect This, in fact, is what Landy and Streightoff found.

The scheme developed in FIGURE 4 may be expected to apply to all species of bacteria in a general way, though the details will certainly vary. The data were obtained by working with a strain of *E. coli* which grows well in a medium of inorganic salts and glucose. Many likely substances were tested for activity by adding them individually to the medium. Only one was found to be active, methionine, and it was then added to the basal medium and all the likely substances tested again, with the result that several more showed activity. It would be of interest to perform the same type of experiment with other species which can be grown in simple, well-defined media

The role of peptones as antagonists has been a prominent once. Since they can be dealt with from the point of view of FIGURE 4, and since the pertinent facts concerning them are listed on page 511, they require little special discussion at this point. It will be recalled that peptone contains a number of antagonists and one dynamist, that of the former PAB is of little importance, and that one (?) unknown agent has a special action upon the heterocyclic derivatives

In experiments comparing *in vivo* with *in vitro* determinations of relative drug potency, peptones are of some special interest. Marshall and coworkers[86] determined the relative potency of sulfanilamide, sulfapyridine, and diaminosulfone against a beta-hemolytic streptococcus, using their *median survival blood concentration* technic. In the mouse, the activities were respectively 1, 1 09 (1.4), and 2 9; in the test tube with peptone broth, 1, 2 9, and 9 3. The figure in parentheses uses the data of Fisher and coworkers[79] to correct the reported ratio for the drug bound to plasma protein. In a later paper on a virulent strain of *E. coli*, the same group[34] showed the ratios to be: SA, 1; SP, 6.3 (8.2); ST, 10.3 (24); SD, 11.2 (12.2). *In vitro* determinations in certain

[86] Marshall, E. K., Jr., Litchfield, J. T., & White, H. J. Jour. Pharmacol. 69: 89 1940.

media showed ST and SD were about 100 times as active as SA, but in the presence of tryptose peptone or of PAB (2 mg. per cent) they were only 8 to 16 times as active. The authors concluded that there "appears to be at least qualitative agreement between *in vivo* and *in vitro* activity of these drugs." To the reviewer, the agreement in the case of *E. coli* seems striking, and it is suggested that the second group of peptone antagonists, defined by the plateau in FIGURE 2, function *in vivo*. The discovery of a specific agent against these would do much to increase the potency of the heterocyclic sulfonamides in clinical practice

The role of nicotinic acid as an antagonist must be considered from two points of view, each of which is different from that discussed above. In the first place, the data of Dorfman and associates[54, 55] establish that under rather special conditions SP and ST can prevent nicotinic acid from acting in the respiratory system of the dysentery bacillus. The presumption is that the heterocyclic rings compete for the same metabolic locus. Granting this, the question arises as to whether such a mechanism is of any importance under the usual conditions of therapy. To the reviewer, the answer would seem to be no, except in those cases where very high sulfonamide concentrations are attained, though it must be admitted that the data available do not permit an absolute decision.

In the second place, growth experiments with the *Staphylococcus*, *Lactobacillus* and dysentery bacillus, for all of which nicotinic acid (or amide) is a growth factor, have shown that, within a limited range of sulfonamide concentration, increasing the nicotinic acid concentration has some antagonistic action Although there is some disagreement in the results reported, and although suitable controls have not always been made, (see page 516 for citations and data), it would appear that the effect is a real one The data do not allow the conclusion that only the heterocyclic ring of the substituted sulfonamide can compete with the pyridine ring, and Wood and Austrian[52] have shown that cozymase can antagonize the quite unrelated thionine as well as any of the sulfonamides. Attention can be focused on the fundamental question by forgetting the coincidence of pyridine rings in sulfapyridine and nicotinic acid. In the case of an organism requiring any specific growth factor, will the concentration of this factor influence the sulfonamide concentration required to inhibit growth (by, e.g., 50 per cent)? Because the metabolic relationships within the cell must be far reaching and complex, it is very easy to believe that this would be so, at least in many if not all cases. Furthermore, it is essential to establish in every case whether the growth factor is present in suboptimal quantities, so

that growth is still proportional to its concentration, or whether an excess is present. There are many species and metabolic types of bacteria, and there are many growth factors. One suspects that many interesting relationships might be found.

In conclusion, I would like to express my regret that this review could not be written with my colleague Major J. S Harris, M.C , who is now on active duty

THE ACTION OF SULFONAMIDES IN THE BODY

By J. S. Lockwood

From the Harrison Department of Surgical Research School of Medicine, University of Pennsylvania, Philadelphia, Pa.

The study of the mechanism of the action of the sulfonamides has been a conspicuously fruitful field for scientific investigation during recent years. Undoubtedly, the most significant milestone of progress in this field was the discovery by D. D. Woods of the competitive relationship between sulfanilamide and *p*-aminobenzoic acid.[1] This was, in fact, the first step in the identification of the latter compound as a substance of general biological importance. At the same time, a strong impetus has been given to the study of competition, in biological reactions, between essential metabolites, or growth factors, and compounds of similar chemical configuration but without corresponding biological activity. It is possible that a clearer understanding of these reactions, coupled with further clarification of the metabolism of cell growth, may lead to progress in a field far removed from that of infection. Of course the immediate result of this work of greatest consequence has been to place sulfonamide therapy of bacterial infections upon a rational, rather than an empirical basis, and the preceding papers in this conference have amply demonstrated the rapidity and effectiveness with which the technics of chemistry and the physical sciences have been utilized to aid the clinician. In fact, these collaborators have carried the study of the mode of action of sulfonamides so deeply into the remote fastnesses of their special fields as to make difficult the task of the clinician in comprehending the results intelligently, and applying all of the implications of their work. It is, therefore, with a sense of inadequateness, but mixed with some pride, that I undertake, as a clinician, to discuss the topic "The Action of Sulfonamides in the Body." The clinician feels inadequate because of the comparative paucity of exact knowledge in the biological application of chemistry, and yet takes pride in the realization that clinicians first conceived the nature of sulfonamide action in general terms, and pointed out the avenues along which fruitful investigation has been

[1] Woods, D. D. Brit. Jour. Exp. Path. 21: 74. 1940.

pursued. For it was Leonard Colebrook[2] who first demonstrated the probable induction by sulfonamide therapy of bacteriostatic conditions in the body, and it was another group of clinicians[3] who first began to relate the effectiveness of the drug to the nutritional environment of the bacteria *in vivo*.

The task assigned to me in this conference involves an attempt to cross the chasm that always separates the test tube from the tissues and to show, as far as one may, that the phenomena presented from the laboratory are important, not simply as exercises in abstract science, but as a rational basis for chemotherapy against bacterial infection. My method of approach will be to point out a few bridges, some of them still incomplete and hazardous to the heavy-footed traveler, by which access to the terrain on the other side of the chasm may be gained. The first bridge is that of "clinical experience": the body of deductions, many of them vague, that any inquiring physician might assemble after applying sulfonamide treatment to a variety of infectious diseases The second bridge, which does not yet completely cross the chasm, is composed of data derived from a study of the reactions of sulfonamides on bacteria that are growing in natural fluids withdrawn from the body, such as blood, serum, and inflammatory exudates, media having compositions that are frequently difficult to characterize but which nevertheless provide conditions closer to the living tissues than can be reached in laboratory media of known composition. The third bridge is provided by microscopic study of tissues and fluids withdrawn from the scene of an encounter between bacteria and host, in which sulfonamide is employed as one of the host's weapons. The final bridge, and perhaps the strongest one of all, affords the opportunity of determining directly in the body whether factors that greatly modify the bacteriostatic action of sulfonamide *in vitro* will similarly modify its action *in vivo* In time, other bridges will be erected, and those already in sight will be strengthened, but ready passage will not be provided until much more is known of the metabolism of bacteria as parasites within an animal host, and of the chemistry of inflammation itself.

CONTRIBUTION OF THERAPEUTIC EXPERIENCE

Careful study of patients led to the conclusion that *the clinical and pathological character of the lesion is a factor of dominant importance in determining its response to sulfonamide therapy.*[3] The characteris-

[2] Colebrook, L., Buttle, G. A. H., & O'Meara, R. A. Q. Lancet 2: 1323. 1936.
[3] Lockwood, J. S., Coburn, A. F., & Stokinger, H. E. Jour. Am. Med. Assoc. 111: 2259. 1938.

tics of completely susceptible lesions are best illustrated in pneumococcic pneumonia as follows:

1. Acute onset.
2. Rapid invasion of the tissue or organ primarily involved by proliferating bacteria. This process might seem to imply a capacity of the bacteria to multiply rapidly in the intracellular fluids without extensive preliminary enzymatic conversion of the substrate. In pneumonia, a period of a few hours may suffice for the lesion to progress from its starting nidus to involvement of both lungs.
3. Maintenance of circulation through the infected area, so that bacteria are frequently found in the circulating blood, where they may or may not continue to multiply. Thus the appearance of positive blood cultures is frequent in many sulfonamide-susceptible diseases and irreversible damage to tissue architecture is not a dominating element in the picture.

Other infections in the same general category include: hemolytic streptococcic cellulitis, meningitis, pneumonia and peritonitis; meningococcal meningitis; and gonococcal infections of peritoneum, joints, and lower urinary tract. When an infection with *Staphylococcus aureus* assumes the general clinical and pathological features outlined above, the disease is much more likely to respond to sulfonamide therapy than is the case when the less rapidly progressive (and more common) course takes place.

The present position of sulfonamide therapy has been reached because of the remarkable success with which it has been applied by physicians all over the world in the treatment of these acutely invasive and formerly overwhelming infections, against which the weapons of immunology possessed only limited effectiveness However, even sulfonamide therapy falls short of being the "therapia sterilisans magna" for which Ehrlich was searching An important limiting factor in the clinical results with these drugs is that once tissue destruction has taken place in any area as a result of bacterial activity or ischemia, this portion of the whole process becomes resistant to the curative action of the chemotherapeutic agents. The physician recognizes that sulfonamide-induced recovery in a severe invasive infection is associated with disappearance of the bacteria from the intact tissue, but expects that active residual infection will remain in regions where the disintegration of tissue has commenced. The problem of healing of the localized abscess is complicated not only by the essential sequences of tissue repair, but also by the prolonged survival of bacteria in the purulent exudate. Therefore, in clinical practice, the physician, when con-

fronted with a case of severe invasive infection, starts out by administering large doses of sulfonamide, and if a progressively favorable course ensues, he may be reasonably certain that no extensive focus of localized suppuration is present. If a progressive trend toward recovery does not take place, he will probably be forced to resort to a surgical procedure to drain or remove a necrotic focus of persistent infection.

If an invasive infection could be likened to a forest fire, sulfonamide chemotherapy serves as a dampening rain, which suffices to check the alarming wind-blown spread of the fire in the underbrush, but does not extinguish the burning of the trees that the fire has already overwhelmed. Within a season or two the underbrush is restored, but the scarred trees remain as semi-permanent reminders of the conflagration. To carry the analogy even farther, just as a rainstorm will not put out a smoldering fire in a peat-bog, so will sulfonamide therapy fail by itself to cure most of the deeply entrenched chronic infections such as bacterial endocarditis, tuberculosis, and the other granulomatous diseases.

On the basis of broad generalizations such as these, the investigator who started his inquiry into the mode of action of sulfonamides from the vantage point of therapeutics quite naturally raised the following questions:

1. Is is not more than a coincidence that the drug effect is of greatest magnitude in the several diseases that are characterized by the most active proliferation of bacteria in the body?

2. Does the highly developed adaptation to parasitism displayed by these invasive organisms (which is perhaps the Achilles heel in their vulnerability to sulfonamide action) result from their utilization of one or more specialized nutritive enzyme reactions which is not true of more saprophytic species and strains?

3. Is the rapidity of spread of invasive and sulfonamide-susceptible infections limited in each case by the available supply in the tissue fluids of some specific nutrient factors?

4. What are the nutrient materials which are readily available to bacteria in tissue fluids that will make it possible for them to multiply actively without first effecting breakdown of the tissues?

Complete answers to these questions are not yet available, but they did serve to provide a motivation and a direction of emphasis for study of the action of sulfonamides in body fluids *in vitro*, and therefore lead us now to an inspection of the second bridge.

BODY FLUIDS AS TOOLS OF BACTERIOLOGIC INVESTIGATION

Leonard Colebrook[3] demonstrated in 1936 the remarkable enhancement of the streptococcidal action of whole blood that followed the addition thereto of even low concentrations of sulfanilamide. It made no difference whether the drug was added to the blood before or after withdrawal of the blood from the body. There was some speculation, at first, as to the significance of the role of leucocytes in this enhancement of streptococcidal action, but Colebrook observed a measure of bacteriostatic action in cell-free serum, and it later became apparent that participation of leukocytes was a secondary, associated phenomenon, and not primarily connected with the drug effect.[4,5]

Since it was known that invasive hemolytic streptococci would normally multiply with great rapidity in the blood of patients dying of an overwhelming infection by this organism, the observance of a reversal of this state after sulfonamide therapy provided a fairly direct characterization, in nonspecific terms, of the nature of the chemotherapeutic action of the drug. Of all the chemical compounds with demonstrable actions against bacteria *in vitro,* only sulfanilamide and its later N^1substituted derivatives appear to retain antibacterial action in the presence of body proteins *in vivo* and *in vitro,* and also satisfy the other basic requirement of a chemotherapeutic agent, that of being absorbed and transported to the site of infection without undergoing chemical inactivation in the process. The real investigation of the mode of action of sulfonamides during the past five years has therefore been directed by students of bacterial metabolism toward explaining this phenomenon in specific biochemical terms. Only knowledge of the biochemical nature of the reaction between drug, bacteria, and host could provide a basis for a rational approach to improvements in chemotherapeutic agents that has been so earnestly desired. However, continuing investigation of the nature of the action of sulfonamides on bacteria in body fluid media has gone hand in hand with these controlled studies of the effects of the drug on bacterial metabolism and has helped to keep investigations of the latter variety from wandering too far afield of the main problem: "How do these drugs work *in the body?*" At the same time, the validity of studies *in vitro* has been supported by the observation of White, Bratton, Litchfield and

[4] Lockwood, J. S. Jour. Immunology 35: 155. 1938.
[5] Lynch, H. M., & Lockwood, J. S. Jour. Immunology 42: 435. 1941.

Marshall,[6] that activity *in vivo* is always accounted for by parallel activity of the drug, or its immediate derivation, *in vitro*.

The following observations seem to go partway toward answering the questions which were raised by "therapeutic experience." When small numbers of *actively multiplying*, potentially invasive streptococci are added to human serum they continue to multiply logarithmically with a generation time of 20–30 minutes for about 10 or 12 hours, after which, if the medium is not supplemented or replenished, the population tends progressively to deteriorate (FIGURE 1). When sulfanila-

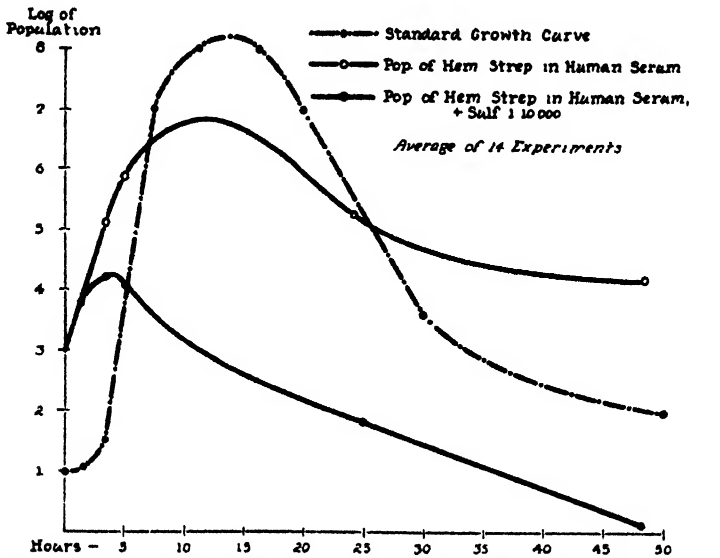


FIGURE 1. Population curve of hemolytic streptococci in human serum, compared to standard growth curve of bacteria (Zinsser)

mide is present in concentration of 10 mg per cent or more, the logarithmic phase continues for only 1 to 2 hours, and the deterioration of numbers of living bacteria commences at about 3 hours. It is true, in general, that the shorter the generation time of growth in control tubes, the more striking is the sulfonamide-induced reversal of the curve in the experimental tubes. In spite of the recognized differences between serum inside and outside the body, there is a general parallelism be-

[6] White, H. J., Bratton, A. C., Litchfield, J. T., Jr., & Marshall, E. K., Jr. Jour. Pharm. and Exp. Therap. 72: 112. 1941.

tween the phenomena demonstrable in the test tube and the phenomena which by implication seem to occur in the body during invasive infection.

To obtain the critical break in the logarithmic phase requires starting with organisms adapted to rapid multiplication in the serum, a medium in which avirulent variants multiply feebly, if at all. Organisms present in media in which they do not multiply are not influenced by "chemotherapeutic" concentrations of sulfonamide[7]

If there is added to the serum a small amount of acid hydrolyzed casein, the rate of growth of the organisms in sulfonamide-free serum is slightly accelerated, and the population tends to reach higher levels before deterioration commences. However, amino-acid mixture does not modify the critical effect of sulfonamide on the logarithmic phase of growth (FIGURE 2).

Addition to the serum of a small amount of enzymatic digest of

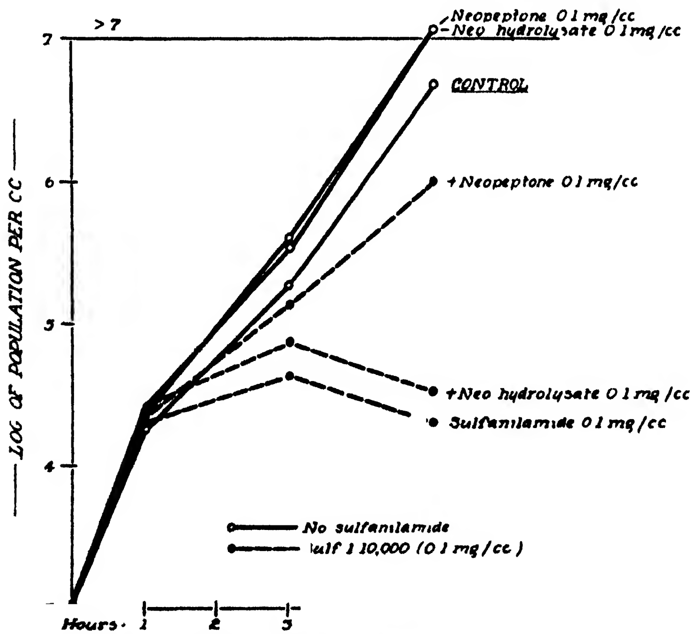


FIGURE 2 Effect of "neopeptone" before and after acid hydrolysis on bacterial growth in sulfonamide action in human serum.

[7] Wolff, L. K., & Julius, H. W. Ann. Inst. Pasteur 62: 616. 1939.

casein (peptone) effects a similar acceleration of growth in the control, affords postponement for many hours of deterioration of the population, and at the same time greatly modifies the reaction of the organisms to sulfonamide (FIGURES 2 and 3). Instead of showing a critical

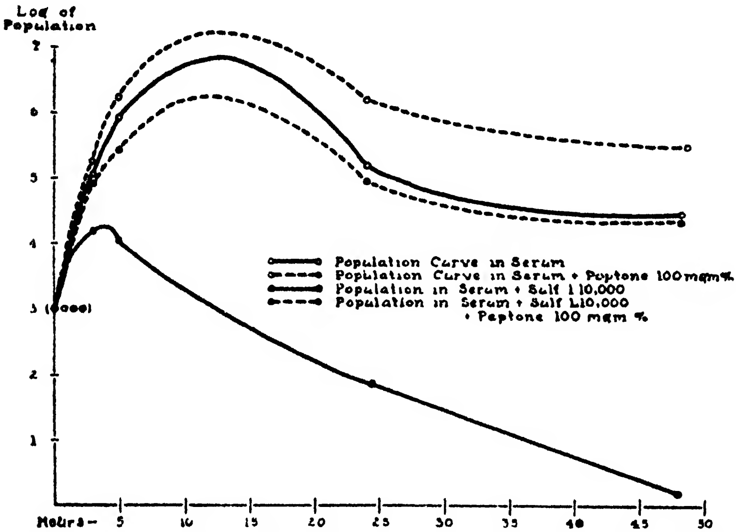


FIGURE 3 Population curves of hemolytic streptococci in human serum Effect of peptone (average of 14 experiments)

break in the logarithmic phase at the 3-hour point, the population continues to increase to a level only slightly below that of the control, and the tendency toward sterilization of the culture is not displayed, as is true in the absence of peptone. Such an experiment appears to reproduce *in vitro* the failure of sulfonamide to influence the survival of bacteria in areas of tissue devitalization *in vivo*. MacLeod has demonstrated the existence of sulfonamide inhibitors in enzymatic digests from a variety of tissues.[8]

Addition to the serum of *p*-aminobenzoic acid causes no alteration in the character of the control population curve, but duplicates the peptone effect in respect to modifying the reaction of the bacteria to sulfanilamide (FIGURE 4). By adjusting the relative concentrations of PABA and sulfonamide it is possible to demonstrate an approximation

[8] MacLeod, C. M. Jour. Exp. Med. 72: 217. 1940.

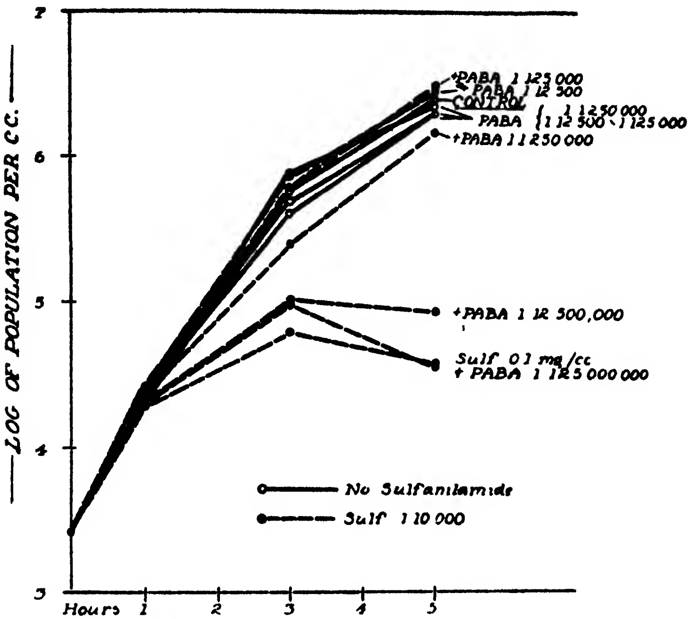


FIGURE 4 Inhibitory effect of PABA on action of sulfanilamide on growth of *Staphylococci* in human serum

of a molar relationship between these reagents. This experiment demonstrates that the competition between PABA and sulfonamide, which is the basis of Woods' hypothesis, holds true in human serum in *in vitro* as well as in laboratory media, an important step toward establishing the validity of this hypothesis.

From these experiments we may infer that sulfonamide lowers the population ceiling of actively proliferating bacteria in unmodified serum, but loses this effect in the presence of certain products of protein degradation, of which *p*-aminobenzoic acid is the most notable one yet characterized. There is reason to suppose that bacteria proliferating rapidly in serum may satisfy their nitrogen requirements by utilization of the limited amounts of nonprotein nitrogen available, that even provided with adequate utilizable nitrogen the completion of an enzyme reaction involving PABA is essential to continued proliferation. PABA may be derived from: (a) digestion of tissue protein, (b) disintegration of bacteria, and (c) artificially added increments.

CHEMOTHERAPEUTIC ACTION AND IMMUNITY

The next bridge must carry the weight required to relate the phenomena of selective bacteriostasis *in vivo* to the modification in the host's reaction to infection induced by the drug. One pier of this bridge consists in the comparison of the microscopic picture of an infection in the non-treated subject with that in the treated subject. Levaditi,[9] and Long, Bliss and Feinstone,[10] independently observed that virulent bacteria inoculated into the peritoneal cavity of the untreated mouse tend rapidly to proliferate, and to overwhelm the phagocytes in the exudate, whereas, in the treated animal, the pullulation of the organisms continues for only a brief period, so that effective clearance of the bacteria is carried out by the phagocytes. Adolph and the writer[11] noted little difference in the type of cellular exudate in untreated and treated rats with induced streptococcic meningitis, but saw a picture of unrestrained proliferation and spread of bacteria in the control animals, contrasting with the appearance of minimal numbers of streptococci in the treated subjects (PL. 4, FIGS. 5 and 6). Only in localized abscesses that sometimes developed in the treated animals was it possible to discern large numbers of bacteria. W. B. Wood[12] studied experimental pneumococcic pneumonia in the rat and verified the observation of drug-induced bacteriostasis in this lesion. He showed, further, that the effect of sulfonamide on the pneumonia lesion differed quite remarkably from the effect of antibacterial serum. The serum promoted active agglutination and phagocytosis of proliferating pneumococci, while the sulfonamide primarily limited proliferation through a direct bacteriostatic action. Organisms at the periphery were "swollen, pleomorphic, and irregularly stained." The same morphologic changes are observed in bacteriostatic studies *in vitro*.[4]

The second pier in this bridge is based on studies of the modification in the development of the immune reactions between treated and non-treated individuals. If the drug effect consists only in the induction of a reduced opportunity for bacterial multiplication *in vivo*, then one might expect recovery from invasive infection to take place without actual development of antibacterial immunity It has, in fact, been observed in both animals and man that drug-induced recovery is not necessarily associated with acquisition of the same degree of immunity that would be exhibited in recovery without drug treatment. McIn-

---

[9] Levaditi, C. Monographie de l'Institut Alfred-Fournier 6. 1937.
[10] Long, P. H., Bliss, E. A., & Feinstone, W. H. Jour. Am. Med. Assoc. 112: 115. 1939.
[11] Adolph, Paul E., & Lockwood, John S. Arch. Otolaryngology 27: 535. 1938.
[12] Wood, W. B., Jr. Proc. Soc. Exp. Biol. and Med. 45: 348. 1940.

tosh and Whitby[13] showed that the height of the immune reaction in mice was determined by the size of the infecting dose, which, of course, determined the amount of bacterial antigen that was absorbed. Where drug treatment was commenced early, and in large doses, very little permanent protection against the specific infection developed. However, where absorption of antigen was increased by either a brief delay in institution of chemotherapy, or by increasing the size of the infecting dose of bacteria, than a higher antibacterial immunity was acquired by the recovering animal. Of especial interest in this connection is the effect of sulfonamide on rheumatic fever, a disease now widely believed to be a delayed consequence of a modified or partial immunity to mild hemolytic streptococcic infection of the respiratory tract. Coburn and Moore[14] and Thomas and France,[15] independently, have shown that if hemolytic streptococcic pharyngitis is prevented by continuous administration of small doses of sulfonamide to rheumatic-susceptible children, then the syndrome of rheumatic fever is not released, presumably because the necessary antigen is not prepared in the upper respiratory tract. However, once a mild upper respiratory infection develops, the administration of sulfonamide not only fails to prevent rheumatic fever, but may actually aggravate the severity of the disease when it develops. This consequence of sulfonamide therapy can be explained on the basis of Coburn's view[16] that rheumatic fever is most likely to occur when the full release of the normal immune reaction to hemolytic streptococcic infection fails to take place at the time of the primary disease By "aborting" the infection the drug actually paves the way for the rheumatic recrudescence.

Sulfonamide therapy is of little value in modifying any of the manifestations of toxemia in infections where the bacteria produce soluble exotoxins. For example, the rash of scarlet fever, and the other phenomena of the acute toxic phase of this disease are not affected by sulfonamide therapy, although the incidence of complications that result from bacterial invasion *per se* may be significantly reduced Therefore, both histological and immunological evidence support the view that the action of sulfonamide in the body is identical with that observed in body fluids *in vitro*, namely, a bacteriostasis that may vary greatly in degree under different conditions, and that is supplemented by other host reactions, but which is qualitatively similar for all drugs and for all types of bacteria.

---

[13] McIntosh, J., & Whitby, L. E. H. Lancet 1: 481. 1939.
[14] Coburn, A. F., & Moore, L. V. Jour. Clin. Invest. 18: 147. 1939.
[15] Thomas, C. B., & France, R. Bull. Johns Hopkins Hospital 64: 67. 1939.
[16] Coburn, A. F. Trans. and Stud., College of Physicians of Philadelphia 8: 91. 1940.

SULFONAMIDE PABA COMPETITION *IN VIVO*

The fourth and final bridge that time permits me to describe to you was put in place securely and well by Selbie,[17] a colleague of D. D. Woods. Immediately following Woods' demonstration of competition between PABA and sulfonamide, Selbie administered PABA to infected animals and showed that the protective chemotherapeutic action of sulfonamide was lost in these animals. It has only rarely been possible to obtain a more clear-cut example than this of identiy of reactions *in vitro* and *in vivo*. PABA did not by itself increase the virulence of the infection, suggesting that the available supply of this substance was not by itself a limiting factor in the rate of proliferation of streptococci in the body, as is also true *in vitro*, but nevertheless in the presence of an excess of PABA the sulfonamide could not adequately block the completion of the specific enzyme reaction that sulfonamide does block when the PABA supply is within physiological limits.

PROMISING AVENUES OF INVESTIGATION

Before concluding this subject, it might be appropriate to summarize the avenues along which further improvements in chemotherapy might be effected. There is, first of all, the possibility of developing new compounds that will compete with PABA in more favorable ratio than do the drugs now available. If this competition is demonstrably augmented not only in synthetic media *in vitro*, but also in serum and other body fluids, and if the drug is transported to the tissues in effective concentration, then even better therapeutic results might be obtained. However, there is as yet little reason to hope for any very important advance along this specific line. More promising perhaps will be attempts to modify the pharmacological properties of the drugs already available, so as to direct their accumulation *in vivo* at the site of active infection in certain types of sulfonamide-resistant diseases. One example of this type of advance is in succinylsulfathiazole,[18] a drug which in large doses resists absorption by the upper intestinal tract, but which in the lower bowel breaks down to liberate sulfathiazole in concentrations that are effective against certain intestinal pathogens. It is a long step from this to the development of a drug that will selectively accumulate in tubercles, or in endocardial vegetations, but perhaps the greatest advances toward chemotherapy of tuberculosis and

[17] Selbie, F. R. Brit. Jour. Exp. Path. 21: 90. 1940.
[18] Poth, E. J., & Knotts, F. L. Proc. Soc. Exp. Biol. and Med. 48: 129. 1941.

bacterial endocarditis may be achieved through such an approach. Shannon's studies[19] are of especial interest in this connection.

Other avenues of progress, which are just beginning to open up, may lead to increases in the activity of the drugs already available. One is the discovery of agents that will selectively combine with or inactivate the substances that now serve as sulfonamide inhibitors For example, if we could find an agent that, under the conditions of Selbie's experiment, would destroy the reactivity of PABA *in vivo*, then a major advance might result. Only in local chemotherapy is there so far much encouragement of practical application of this principle. It has already been suggested that urea,[20] azochloramide,[21] and certain oxidizing agents are capable of synergizing sulfonamide action by an inactivating effect on PABA and its precursors. The control of pH as a means of increasing the activity of sulfonamide through promotion of ionization[22, 23] is also impractical in systematic chemotherapy but may prove to be useful in treatment of local surface infection if suitable buffers can be developed. How far it will be desirable to go in exploring these avenues for increased sulfonamide activity will depend in part on what limits appear in the chemotherapeutic possibilities of the newer antibiotic agents such as penicillin and streptothricin. The brilliant success with which penicillin is now being used experimentally in the treatment of invasive infection by staphylococci and other sulfonamide-resistant bacteria leads us to believe that the broadest highway for immediate progress is through this field. However, so unique is our understanding of the mechanism of action of sulfonamides that there is a challenging opportunity for bold application of this fundamental knowledge toward the ultimate attainment of Ehrlich's chemotherapeutical ideal. We have attempted to show that tested experimental tools are already available by which the antibacterial activity of intelligently designed sulfonamide compounds may provide a basis for interpreting chemotherapeutic reactions in the body.

As Galdston[24] has recently pointed out, it is interesting to note, that our present concept of the action of sulfonamides in the body is in close accord with an early hypothesis advanced by Ehrlich in 1908 to explain recovery from infection, that of "atreptic" immunity: ". . . it is sufficient to assume that those substances (bacterial nutrients) may

[19] Shannon, J. A. Ann. N. Y. Acad. Sci. 44: 455. 1943.
[20] Clark, W. G., Strakosch, E. A., & Nordlum, O. Proc. Soc. Exp. Biol. and Med. 50: 43 1942.
[21] Neter, E. Proc. Soc. Exp. Biol. and Med. 47: 303. 1941.
[22] Fox, C. L., & Rose, H. M. Proc. Soc. Exp. Biol. and Med. 50: 142. 1942.
[23] Schmelkes, F. C., Wyss, O., Marks, H. C., Ludwig, B. J., & Strandskov, F. B. Proc. Soc Exp. Biol. and Med. 50: 145. 1942.
[24] Galdston, Iago. "Behind the sulfa drugs." Appelton Century Co New York, N. Y. 1943.

still be present (in the body), but that the parasitic agents in question are incapable of absorbing them; in other words, that the substances have ceased to be at the disposal of the bacteria" . . . "It suffices to suppose that the bacteria . do not find the needful means of existence in the body and therefore cannot multiply. This being the case, they cannot for any length of time remain alive in the body, for then the latter's defensive forces, its phagocytes, come into action and destroy the invaders in a non-specific manner" A theory to account for spontaneous recovery from infection, which was at one time propounded by Ehrlich but later abandoned, may now be extended to describe with remarkable accuracy the probable mechanism of the chemotherapeutic action of the sulfonamides *in vivo*

PLATE 4

Meningeal exudate of rats experimentally infected with *Streptococcus hemolyticus* (see Page 534)

FIGURE 5 Exudate of control animal dying at 40 hours containing large masses of free cocci

FIGURE 6 Exudate from sulfanilamide-treated animal sacrificed at 48 hours No organisms visible

(For complete details see Adolph, P E and Lockwood, J S, Arch Otolaryngol, 27:535, 1938)

FIGURE 5

FIGURE 6

Annals of The New York Academy of Sciences
Volume XLIV, Art. 6. Pages 539-624
December 22, 1943

PSYCHOSOMATIC DISTURBANCES IN RELATION TO PERSONNEL SELECTION*

By

Lawrence K. Frank, M. R. Harrower-Erickson, Lawrence S. Kubie, Gardner Murphy, Donal Sheehan, and Harold G. Wolff

CONTENTS

*This series of papers is the result of a conference on Psychosomatic Disturbances in Relation to Personnel Selection held by the Section of Psychology of The New York Academy of Sciences, February 5 and 6, 1943. The Academy is very grateful to Doctor Anne Roe for assembling the papers presented at this conference and for rendering very able editorial assistance.

Publication made possible through a grant from the Conference Revolving Fund.

INTRODUCTION TO THE CONFERENCE ON PSYCHOSOMATIC DISTURBANCES IN RELATION TO PERSONNEL SELECTION

By Lawrence K. Frank
New York, N. Y.

This conference on Psychosomatic Disturbances in Relation to Personnel Selection has been called to focus attention upon some of the urgent problems of personnel selection for the war and the postwar period.

Under necessity of mobilizing our total man power and woman power for the war effort, in the armed forces, in industry, agriculture and essential civilian activities, we are faced as never before with the question of how to select personnel for the various tasks involved, in such a way as to conserve our human resources and to avoid assignments of individuals to tasks that will unnecessarily expose them to breakdowns or damage The casualties of combat will undoubtedly be large, but we are increasing our human wastage by unwise, careless selection and assignment of personnel, and by disregard of individual differences

We are far from having the methods and procedures that will unerringly select individuals for each of the many activities of wartime according to their individual capacities and peculiar fitness for such work At best, today, we can attempt only the more modest task of trying to select the individuals who should *not* be assigned to various activities for which they are not suitable But this modest endeavor, if carried out with the knowledge and resources now available, might prevent much unnecessary wastage of human lives incident to total war and at the same time enhance the contributions which individuals, each according to his or her capacity, might make.

The problem of selection of personnel will not be lessened when the war ends. Rather we shall face an enormous undertaking in demobilization, rehabilitating and retraining the men and women in the armed forces, in war industries and in the other activities that will be rapidly curtailed when fighting stops. It has been proposed that a systematic effort be made to assess each individual to be demobilized for possible further education, vocational or professional training and for job placement. This presents a challenging opportunity to go beyond the

present rather limited procedures, many of which rely upon a person's past experience, what he knows or has done and his limited acquaintance with certain jobs or the terminology of a profession. We must begin to reveal the potentialities of individuals often unknown to themselves and their levels of aspiration, thereby unlocking large human resources that we are now ignoring or wasting.

For those who have been wounded and maimed, this postwar program will call for more or less extensive rehabilitation, occupational therapy and the restoration, so far as possible, of their ability to take up the tasks of living again despite handicaps, impairments and emotional disturbances. Every injured individual, it seems safe to say, will be in need of psychotherapy.

It seems probable that this war, requiring prolonged periods of waiting under mounting tension before going into action, as in submarines, airplanes, gliders, parachute corps, tanks and the like, will foster a heavy incidence of psychosomatic disorders since individuals will have to "consume their own smoke" and carry their mounting anxiety in their organ systems and functional processes. Thus it seems likely that in the postwar period there will be a heavy load of such patients for which we are scarcely prepared either to diagnose correctly or to treat effectively

Likewise it seems probable that there will be an increasing frequency of disturbance in civilians who will exhibit, either during the war or after, symptoms of dysfunctions, mental disorders, or crippling neuroses We can not expect a people, who have gone through ten years of a major depression with all that it has done to families and individuals, to have the necessary resilience to accept the wartime controls, deprivations and restrictions with ease Let us remember that we are now asking people to accept deprivations of food and fuel, shortages of customary goods and services, to live in poor and congested housing, use congested transportation, meet rising prices and mounting taxes, etc., with all the current insecurity and worry, especially over sons and husbands in the armed forces, some of whom are already being reported as casualties Unlike the British who are a more homogeneous people, with more intact traditions, confronted with immediate danger and destruction, we may not be able to "take it" so well. It seems likely, therefore, that we shall have a heavy load of civilian cases in addition to those of service men and women after the war, for whom we must prepare needed procedures for diagnoses and therapy.

If we recall these situations and prospects, the full description of which would justify a number of conferences, this meeting appears

highly appropriate and sufficiently important, professionally and socially, to warrant calling upon busy investigators to prepare papers and come here for their discussion. We must make every effort to clarify our problems and improve our procedures for understanding individuals if we are to meet the present and future demands for professional guidance and direction in selecting personnel

Before we turn to these papers and embark upon the discussion of the many technical problems to be considered, it may be permissible to indicate briefly the larger context in which we find ourselves today

The preoccupation of scientific research with a search for regularity, uniformity and unvarying order, to be expressed in laws, generalizations and equations, has had as a consequence the development of a belief that any concern with the individual was, however important for other purposes, not scientific Indeed, in many statements and published papers one gains the impression that the individual who departs from the statistical norm or deviates from the uniform pattern is a curious and rather deplorable breakdown in Nature's otherwise majestic orderliness This search for uniformity has been the pattern of the physical sciences which, as we are now realizing, have until recently focused upon those events that were the outcome of many convergent phenomena that averaged out in more or less regular and uniform order.

The prestige and the success of these methods have greatly influenced the life sciences which have more or less explicitly ignored the individual and concentrated upon a similar search for regularity and uniformity

In much of the work of the life sciences we see how frequently individuals are observed and measured for data to show the relation between two variables, but the individuals who supply such data are of little or no interest and are regarded as having no significance for the specific problem under study. This preoccupation with professional problems has involved using individuals as sources of data, which are then abstracted and treated as wholly independent of the individual organisms from whom they were derived, as if the data were more or less absolute, especially if quantified

Indeed it may be said that in the life sciences we have endeavored to understand individuals primarily by studying groups and age and sex group characteristics For example, in psychological procedures devised for measuring individual differences, the majority of the tests have been developed as ways of discovering how much an individual departs from the norms of action, speech and belief established statistically for each chronological age and sex group. Likewise, in many

structural and functional tests, the individual is viewed as deviating from certain norms of size and performance for age and sex. These tests and procedures are of undoubted value in classifying individuals according to their conformity to, or deviation from, such norms, but it seems fair to say they do not reveal much about individual differences as such nor give much understanding of the individual's idiomatic, if not unique, make-up and functioning, of how he manages to live and carry on with those deviations from the norm Only a few investigators and a few tests have accepted the challenging problem of studying individuality in its manifold expression.

Within recent years the climate of opinion in scientific work has been changing. With the development of quantum physics has come the realization that beneath the uniformity and regularity of events there is much disorder, irregularity and lack of uniformity, as exhibited by the electrons, atoms, molecules and other constituents of the larger aggregates that have been found so orderly and uniform by classical methods. Almost suddenly, therefore, physicists have begun to study individuals, to say explicitly that many of their most cherished generalizations and laws are applicable only to convergent phenomena * Thus we may say that the study of individuals has begun to become scientifically respectable and therefore the students of the life sciences need not be too hesitant to embark upon such investigation, which now offers probably the most exciting and challenging problems of our time.

If we are to undertake the study of individuals as a search for something more significant than the description of mere uniqueness, then we must prepare to revise some of our customary ideas and procedures

It appears that increasingly we shall focus our thinking not only upon the problem of the relation of two variables, studied in multiplicity of cases, but upon the problems arising from an interest in patterns, configurations and organizations. The field concept leads to a reformulation of the idea of "parts" and the "whole" and of the problem of organization. Instead of seeing "parts" as discrete entities or variables that are coerced into an organized whole by some unknown power or force or vitalistic quality, we begin to see the whole as that which the so-called "parts" create by their patterned interreaction which in turn patterns the activity of the parts We can therefore no longer rigidly isolate "parts" from their organization nor can we pick out one organ system or functional activity in the organism and endow it with

*See Langmuir, Irving Science, January 1, 1943.

a potent causal power over the other organ systems or functions, as so much of our thinking in physiology seems to imply. Indeed, we are often told that a certain gland, such as the pituitary, is the cause of certain other organic events, as if the pituitary were somehow outside of the total organic complex and not being acted upon by its presence in the organism. The notion of cause and effect, of a *vis a tergo*, still dominates much of our thinking in the life sciences and by so much blocks our efforts to understand how an organism functions, and how the several organ systems participate in the organic totality. For the life sciences faced with the problem of organization, the field concept offers a release from many traditional problems that derive from the older concepts of organization; an organization becomes the totality of interrelations—of actions, reactions, interreactions that take place among the so-called "parts," thereby creating and maintaining the organized whole that in turn patterns the activities of those parts. Organization is not a mysterious power or entity or prime "cause." When we take the concept of organism seriously, we will cease to reify data into independent entities that are then abstracted from their organic field and so give rise to artificial problems.

The field concept means biological relativity and the recognition that every measurement or observation we make must be ordered to the field—that is, studied and interpreted in terms of the organism in which it occurs and has been observed. Specifically this means, as I interpret it, that the same magnitude observed in a number of organisms may have widely different significance while different magnitudes will have the same significance, because each magnitude or other observed characteristic must be enlarged or reduced according to the organism exhibiting that magnitude. This, I take it, is what the clinical method does, translating every finding into its relative significance for the total individual patient under study and treatment.

It likewise indicates that we should shift our focus from a search for identities and quantitative uniformities to a study of equivalents, recognizing that it is the configuration, pattern or organization that is significant, not merely the isolated parts or dimensions, which may vary widely within the same configuration. In other words, the future trend in the life sciences will probably be to study organization and processes, not merely end products, since the same process may give rise to a wide variety of products which we must learn to recognize as equivalents coming from the same process. This shift is already recognizable in some disciplines, such as genetics, where the most dissimilar structures or organisms are being revealed as genetically related, that is,

coming from the same basic process. The criteria of credibility plays a large role in scientific advances, as we see in physics today, where the criteria for large-scale convergent phenomena are being superseded by other criteria for quantum physics Likewise, new criteria for study of individuals are needed since coefficients of correlation, measures of dispersion, sigma distributions are not very applicable. Nor is factor analysis the answer, as that also necessitates fractionating individuals into discrete measurements that can be put together only for the group.

This does not mean scientific chaos or irresponsible hunches and disregard for precise quantitative procedures. On the contrary, it calls for much more refined and rigorous methods and the manipulation and interpretation of data by more appropriate concepts and procedures, just as the change from classical physics to quantum physics has brought, not confusion and retrogression, as some feared, but great advances and more understanding.

If we also take seriously the concept of the four-dimensional space-time, we shall see that such ancient dichotomies as that of structure versus function become not only useless but obstructive * The problems that the dichotomy of structure versus function have created can be resolved by showing they are not real problems but artificial problems generated by unreal assumptions †

For example, much of resistance to the psychosomatic viewpoint and approach to individuals arises from the persistence of the older notions of a rigid structure and of textbook functional norms and the concept of a cause and effect relationship supposed to govern all organic reactions These were abstractions that are no longer credible nor useful but are so strongly entrenched in our thinking that we refuse to look at the actual structural picture a plastic, ever-changing scheme of spatial relationships Likewise, we are reluctant to accept the available data which show that organisms rarely conform to the simplified textbook norms or artifacts Moreover, as Dr Sheehan has emphasized, every individual has his or her idiomatic pattern of homeostasis with a greater or less variability from period to period (variability being the change within the organism as contrasted with the variation from organism to organism) Here in homeostasis we see an excellent illustration of how the same process may appear in different magnitudes and give rise to different or similar products -the stability of the individual which he maintains at a certain physiological cost

The recent studies with tagged atoms and isotopes are showing that

* Frank, Lawrence K Structure, function and growth Phil of Science 2: No 2 April, 1935.
† As Chancellor Kemp Smith remarked many years ago, "The history of human intelligence is a record, not so much of the progressive discovery of truth as of our gradual emancipation from error"

the organism is in a continual state of flux, as the environmental impact (weather, food, etc.) acts upon the organism, which selects what it will react to and what it will retain or release. Even teeth are constantly being replaced, atom by atom.

This is not the time to enlarge further upon the many far-reaching changes that impend in the life sciences. As I have on previous occasions suggested, we are seeing the rise of a biological relativity which will open as rich and promising a new field of work as did physical relativity.

Once we have developed a consistent and thoroughgoing organismic concept and accepted biological relativity, we shall cease to be troubled by such questions as, "How can emotions *cause* a disturbance in the cardiovascular system or gastrointestinal tract?" We shall realize that the question is no longer valid but that we can and must proceed to study how organisms react under varying threats and loads and deprivations, including the persistent threats carried in the organisms as residues (physiological changes) from previous experience.

We shall also begin to focus on the question of organic incongruity and discrepancy, studying the interrelationship among the often imbalanced organ-systems, structures and activities in the individual and seeing his idiosyncratic behavior and functioning as arising from his peculiar configuration in his multidimensional environment.*

It may also be pointed out that we have not yet clarified the idea of time in the life sciences. Time, as I see it, is more than a variable or a dimension in a frame of reference for ordering observations. Time is a process as we realize when we reflect upon biological or physiological time. Moreover, we must try to clarify the concept of past experience and give it a formulation compatible with our other new concepts and orientations. So much perplexity in this field of study arises from the confusion over past experience and inability to see that the organism continues to carry its past experience into the present as persisting modifications of the organism.†

What I have said on this subject is appropriate, I believe, because the psychosomatic concept is at once a protest against the older fractionation of the organism into supposedly discrete parts and separate powers and *ad hoc* causes invoked to explain every functional activity and behavior, and it is also a program for research and therapy directed toward a more holistic or organismic field concept of man. Indeed it is not unwarranted to say that in psychosomatic research we

* See Man's multidimensional environment. Scientific Monthly, 56: 344-357. April, 1943.
† See Locus of past experience. Journal of Philosophy 20: 327-329. 1923.

are seeing the forward movement of the life sciences, some disciplines of which have scarcely realized the significance of this new climate of opinion.

There is another aspect of this conference that I cannot forebear discussing at least briefly. The central issue of our time, dramatically and tragically focused by the war, is the value and significance of the individual human personality. Over and above voting and representative government, of freedom of action, speech and belief, is the crucial question of freedom for the human personality. Now there is no one more keenly aware of how we are neglecting and mistreating the human personality than those who as physicians, psychiatrists, psychologists and others are studying, treating and attempting to rehabilitate the immense army of men, women and children who are warped, twisted and distorted, driven by anxiety, hostility and guilt, exhibiting the many forms of mental disorders, of illness, of neurotic defeat and socially destructive conduct

If we are really convinced that the democratic way of life is important and must be protected and more fully realized, then I submit that we must strive to understand individuals and learn how to rear children, educate youth and employ men and women so they can live at peace with themselves and therefore at peace with society It cannot be too strongly emphasized that if we desire a free, democratic social order, we must protect and develop each individual so that he is capable of carrying the burdens of freedom, of helping to maintain social order by self-disciplined co-operative conduct and an awareness of, and respect for, the personalities of others We cannot therefore permit anyone, no matter how lowly, insignificant or seemingly unimportant, to be humiliated, degraded, terrorized or otherwise damaged as a personality, because those so treated are unable to participate in a free democratic society or to develop the kind of human relations to which all aspire

Specifically, if in our homes, schools, clinics and hospitals, workshops and other places of human contacts and associations where we assail the integrity of individual personalities, we are to that extent denying the democratic aspiration and blocking its attainment The psychosomatic concept is, therefore, a very real and effective expression of the democratic ideal and a protest against all those practices, lay and professional, that depersonalize and deny the respect and the human treatment that each individual merits as a personality.*

Perhaps in the dark and difficult days to come, those who are work-

* See Freedom for the personality. Psychiatry 3: No. 3. August, 1940.

ing on this problem of understanding and selecting individuals and conserving personalities may have to carry the major responsibility for the further development and application of this democratic ideal to life They may be the only group in a position to assert that ideal and to indicate what we must do to reconstruct our traditional culture and our social life to attain that goal A recognition of our present and future responsibilities of this kind may serve to give this meeting and the activities which we hope it will stimulate an added meaning and significance

SOME PHYSIOLOGICAL PRINCIPLES UNDERLYING VARIABILITY OF RESPONSE

By Donal Sheehan

College of Medicine, New York University, New York, N. Y.

To cover in any adequate manner the topic that has been given to me as the subject of my remarks would encompass, and might even end, with an account of the structure and functions of the living cell. To do so would sidetrack me into a discussion of internal metabolic processes, the principles of which are familiar to you, and the controversial details of which I am ill-prepared to debate.

Sherrington has said that in the life of the organism, the cell is "a unit-life dynamically and structurally. Its shape and visible parts, commonly called structure, are indeed in fact as dynamic as any of its other features. To appearance more stable, as boundary membranes, etc., they are none the less steady states or moving equilibria and as 'living' as the rest. The cell is a polyphasic system whose total average dynamic equilibrium rests on energy-exchange between its parts and between them and its surround, an energy-exchange organized so as to centre on itself."* Its activities can, for the most part, be expressed in known physical and chemical processes. The energy is turned to maintaining itself. "To behave in this way" continues Sherrington "is a common and convenient phrase to manifest life. It involves dependence on its surround for energy. It is a conception unthinkable apart from its surround. It is so locked into its surround that to extract it thence is to break it in all directions."

This summarizes aptly the first point which I wish to stress: The essence of life is change. Any reaction of an organism or of its parts to a new stimulus is superimposed upon a base line, which is already fluctuating, and which is a reflection of the organism's response to an ever-changing environment. There is no such condition as "at rest," and this is true no matter what particular type of living cell we may be studying. When we speak of a physiological "constant" of the body temperature being 98.6° F., or of the blood sugar level as 80 mgm. % – we are referring to a mean value from which fluctuations occur in either direction. The "constant internal state" of Claude Bernard is a condition not of static but dynamic balance. It is maintained at a price

* Sherrington, C. S. Man on his nature. Cambridge University Press.

Forces pulling in opposite directions are equilibrated so as to give an appearance of rest. This is nowhere better illustrated than in the standing posture, where lack of movement is only maintained by tonic activity in both the flexor and extensor groups of muscles. The dynamics of the circulatory system offer another example of delicate balancing mechanisms interlocked to maintain an adequate venous return to the heart, with depressor and vasomotor reflexes called into play the instant the blood pressure fluctuates beyond the limits of the physiological needs of the tissues. Innumerable other illustrations could be given.

In the limited time at my disposal, I must focus my discussion on the specialized cell that forms the unit of the nervous system. But I must digress for a moment to indicate the common ground from which the physiologist and psychiatrist can begin discussion. The motor response of the individual is the fabric of both the physiologist and the psychiatrist The physiologist interprets it in terms of neural activity, the psychiatrist in terms of psychological behavior. If there is a true liaison between "psyche" and "brain," one might at least hope that the piecing together of the factors of neural activity would lead to a better understanding of the psychological process. We have to confess that we are as yet far from realizing this expectation Let me quote again from Sherrington.

"But where does neurophysiology contribute anything to the knowledge of the norm from which anxiety causes departure, and what has cerebral physiology to offer on the whole subject of 'anxiety'? The psychiatrist has perforce to go on his way seeking things more germane to what he needs The mind is a something with such manifold variety, such fleeting changes, such countless nuances, such wealth of combinations, such heights and depths of mood, such sweeps of passion, such vistas of imagination, that the bald submission of some electrical potentials recognizable in nerve-centres as correlative to all these may seem to the special student of mind to be almost derisory. It is, further, more than mere lack of corresponding complexity which frustrates the comparison The mental is not examinable as a form of energy. That in brief is the gap which parts psychiatry and physiology. No mere running round the cycle of the 'forms of energy' takes us across that chasm "

This is admittedly the sad truth, though the gap is perhaps less if we do not try to jump across it before we have attempted to construct a bridge. We have already noted that both the physiologist and psychiatrist are essentially dealing with the same material. The physiol-

ogist may unravel the mechanism of nerve conduction, but he does so in order to explain certain manifestations of "pain," or he may record electrically the changes in tension in a skeletal muscle, but thereby hopes to analyze the component parts of "voluntary" movement "Pain" and "voluntary" movement are likewise the problems of the psychiatrist. Fundamentally, I suppose the physiologist asks the "how" and the psychiatrist the "why" But these questions are not so easily separable, and the paths of the physiologist and of the psychiatrist must inevitably cross

There is, to be sure, a difference, not so great perhaps as one would first suppose, in the methods of physiology and psychiatry The psychiatrist is concerned more with the reactions of the organism as a whole The physiologist's approach to the same problem is to reduce the factors of variability to the minimum, to study the functions of the whole in terms of its separate parts The very nature of the experimental setup is artificial So indeed it must be admitted is the psychiatric approach, for the mere presence of the psychiatrist is a factor superimposed on the previous environment Physicians are not aware of this as acutely as they might be Likewise the experimental setup of the physiologist is by no means exact The very presumption of life demands, as we have seen, a fluctuating base line from which the experiment starts If these fluctuations are small in comparison to the artificial responses produced in the experiment, as they are in skeletal muscle, then the relation of the experimental observations to the particular stimulus can be more readily deduced. If, however, as in smooth muscle, the experimentally produced responses are superimposed upon a widely fluctuating base line, interpretation is considerably more difficult, and the physiologist's setup is closer to that of the psychiatrist.

I take it that it is my particular function at this meeting to indicate how the structural pattern of the nervous system and its activities, as studied by the physiologist, are oriented toward variability of response And let me say at once that this indeed is the basis of the physiological approach. The concept of rigid constants is an erroneous one, which finds no support in any physiological work.

Let me begin with the unit of the nervous system—the neuron as it has been called—which consists, as you are well aware, of the nerve cell-body together with all its processes, both dendritic and axonal That it is a structural unit can be demonstrated by the death of all its parts which follows destruction of the cell-body, its so-called "trophic center" Such units are linked together in the central nervous

system in a highly complicated pattern by junctions called synapses, at which, as Cajal tells us, there is "contact without morphological continuity."

Such interfaces are in reality multiple points of contact, the branching axon and its many collaterals terminating as end-knobs or "boutons," 2–5 μ in size, applied to the surface both of the dendrites and of the cell-bodies of other neurons It is estimated that several hundreds of boutons make contact with one neuron, so that the total interface is very large. The physicochemical changes that occur at such interfaces account for alterations in activity in one neuron consequent upon the changing activity in another adjoining neuron, or as we say simply, for the conduction of the nerve impulse Conduction "across" the synapse is believed to take place only in one direction Threshold stimulation of the neuron takes place, as Lorente de Nó has expressed it, whenever all, or, at least, the majority of knobs at a discrete zone of the neuron are activated simultaneously, or within a very short interval of time.

The processes of certain of the neurons within the central nervous system stretch out to the periphery, on the one hand to "receptors" or sense-endings, on the other to "effectors," the muscles and glands On the motor side, it has been shown for the cat that a single motor nerve fiber, by extensive peripheral branching, innervates as many as 150 skeletal muscle cells The innervation ratio varies, so we have small and larger motor units, designed presumably for discrete and more diffuse responses, respectively On the sensory side, the peripherally running processes of neurons terminate in a variety of forms The threshold of response to different modes of stimulation varies in the different morphological types of sensory nerve ending, and specific modalities of body sensation have been assigned to each, sometimes on rather incomplete evidence. As on the motor side, each sensory nerve fiber shows extensive branching in the periphery and its total area of supply may be considerable In the cornea, for example, Tower estimates that a single sensory fiber may innervate an area of several mm diameter. Furthermore, the areas of such individual "sensory units" overlap, one with another, so that stimulation at any one point may produce activity in several nerve fibers Neural activity from a single stimulus is therefore probably always spatially multiple from its inception.

The physiological unit of the nervous system is generally considered to be the reflex. Thomas Willis, who introduced the term, and later Robert Whytt, used it to imply an act where stimulus is promptly

followed by movement without conscious participation of the "will." It is true that many reflex acts can be elicited uniformly in a spinal animal, but differences in the timing or intensity of the stimulus can alter the response, and the reflex act is by no means regular quantitatively. Furthermore, a reflex movement operated from one peripheral stimulus can be modified by a second stimulus applied at a different place, as is shown in the well-known inhibition of the crossed extensor reflex

In the intact animal not deprived of higher neural centers, the variability of reflex response becomes more marked It is familiar to all of us that a "sluggish" knee jerk can be reinforced by the patient voluntarily gripping his hands at each tap of the patellar tendon, and that in the so-called "nervous" individual the stretch reflexes from tendons become hyperexcitable.

The term reflex has been extended by Pavlov and his school to include "conditioned" reflexes developed through habit formation, so that the participation of the cortex becomes an integral part of reflex action in this sense I will leave the discussion of the conditioned reflex to abler hands, and merely indicate here that the unconditioned reflex is also dependent upon antecedent neural activity, although the time interval is of a very much shorter duration

Let us consider for a moment the structure of the spinal reflex arc Some of the afferent neurons entering the spinal cord form synaptic connections directly with the motor horn cells from which issue the axons to the muscles themselves More often, at least one interneuron is placed in the pathway between the afferent and efferent limbs of the arc This is the simplest concept of the reflex, and it may be ipsolateral or contralateral. It does not, however, represent in any way the existing pathways open to an entering afferent volley of impulses. Each afferent fiber divides on entering the dorsal part of the cord into ascending and descending branches, each of which sends off several collaterals. The secondary neurons which are brought into synaptic relationship with it are many, and lie at various levels above and below the point of entry of the primary neuron The arrangement allows for "dispersal" of the effect of the impulse within the central nervous system. The axons of the *secondary* neurons also give off collaterals, so that in turn each secondary neuron comes into synaptic relationship with more than one tertiary neuron, allowing for still greater dispersal, or in some instances "convergence" of the effect of the impulse.

A newer concept of activity within the interneuronal pool has been developed through the researches of Forbes, Lorente de Nó and others It now seems probable that both "delayed" and closed self-reëxciting

site side of the cord and brain stem as far as the thalamus, and from thence to the postcentral cortex. Such diagrammatic representation can be found in almost every textbook of neurology. I do not wish to imply any lack of significance of the certain well-defined tracts which pass up and down the cord and brain stem, but the necessary emphasis on these have, I think, obscured the true morphological pattern of the internuncial neurons. This oversimplification of pathways, and perhaps an unfortunate emphasis on the term "center," have made it difficult rather than easier for physicians to understand many symptoms with which they are constantly confronted, and which seem bizarre and hence "imaginary," solely because of too-restricted concept of the structure and function of the nervous system.

The long conduction pathways are obviously of utmost importance in the mechanisms of sensation and probably act as the principal channels of conduction for more immediate responses. But, if we look back to the multiple points of termination of each entering afferent fiber and the collateral branching of the secondary neurons, what multiplicity of paths are open to each excitatory process! And, if closed self-reëxciting circuits are present, the patterns of excitation reaching the cortex from a single afferent stimulus must be spatially and temporally very different from the single nerve impulse which may have originated them. Indeed, we come to wonder how stimulation of the body surface is so accurately localized. Such spatial localization of sensation is, of course, in the nature of a conditioned response, dependent upon previous experience.

When we turn to mechanisms of autonomic regulation, variability of response becomes the rule. Let me just emphasize one or two points. The same branching of the axon is found in the preganglionic fiber so that it connects with many postganglionic fibers. The basis for a wide and diffuse response is obvious. Secondly, sympathetic fibers probably reach every part of the body, traveling along blood vessels, but the parasympathetic distribution is more restricted, so that not all structures are innervated from both divisions of the autonomic balance. Where both are present, a reflex response in a tissue (smooth muscle or gland) may be due to loading the scale on one side or diminishing the weight on the other, and it is not always easy to distinguish. And, even when all extrinsic nerve supply is removed, smooth muscle retains rhythmic tone. Some of this is due to hormonal influences. Adrenaline, when liberated in quantity, can reproduce with minor modifications the physiological picture of diffuse sympathetic activity. But even after complete sympathectomy, where the adrenals are vir-

tually inactivated by denervation, the cat can maintain its body temperature and blood pressure in the protected existence that a laboratory offers, though it fails to do so when exposed to situations of undue stress requiring rapid adjustments. Thus the organism can call upon one of several mechanisms to bring about the identical response.

On the other side of this picture of variability in autonomic regulation, we find that the same responses may be brought about by two entirely different stimuli. Consider the reaction of a warm-blooded animal on exposure to cold. The processes of heat production can be stepped up, and those of heat loss proportionately diminished. Increase in the activity of skeletal muscle is by far the most important method of heat production in rapid adjustments. The earliest response to cold is an increase in muscular tone (so-called "tensing" of the muscles) and continued exposure to cold leads to shivering Heat loss is diminished by decreasing the rate of blood flow through cutaneous vessels, which can be brought about by increasing the tonic constrictor control of arterioles. These vasomotor responses to changes in environmental temperature are greatest in the extremities and particularly in the digits of the hands and feet The extreme is seen in the dead-white fingers of a Raynaud spasm.

Shivering and vasoconstriction are two of the main reactions of a warm-blooded animal to a fall in environmental temperature But exactly the same response can be brought about by emotional stress, and often in the same individual. A young girl, who was subject to typical Raynaud attacks whenever her hands were exposed to cold, was lying in bed in the hospital ward in a warm environment. On the entry of the surgeon into the ward, the fingers of both her hands went dead white, the muscles in her limbs became tense, and she began to tremble, which is surely the same as shivering, a series of synchronous contractions in groups of muscle fibers, repeated out of phase with those of other groups. This patient showed another interesting feature. She gave a history of major epileptic seizures which were always brought on by some situation of conflict in her home and which were preceded by a typical Raynaud attack. In the course of our investigation we used the intravenous adrenaline test devised by Freeman, Smithwick, and White. Normal saline was run into the median basilic vein at the rate of 40 to 60 drops per minute. After half an hour, an adrenaline solution of 1 to 250,000 saline was substituted, the patient being unaware of the change. Within 5 to 10 minutes, a typical Raynaud attack was precipitated, unknown to the patient, who was

arranged so that she could not see her hands. And then, within a minute or so, she had a genuine *grand mal* seizure.

I mention this case to illustrate how similar responses may follow different kinds of stimulation. Increased skeletal muscle activity together with vasoconstriction formed a recognizable pattern of response to: (1) stimulation of somatic afferent fibers by the application of cold; (2) a situation of "conflict," with consequent emotional stress; and (3) an intravenous injection of adrenaline.

Notice that the responses were evident in both somatic and autonomic spheres. Indeed it is questionable in my mind whether there is ever a purely somatic or a purely visceral reflex. I want to use one other clinical illustration, which shows this interrelationship in an even more striking manner, and which, incidentally, was the first problem that called my attention to the importance of the psychosomatic aspect of medicine. It will, I think, be pertinent to the more immediate discussion of this conference.

During the Civil War, a special study was made of peripheral nerve injuries. Out of this emerged the recognition, by Weir Mitchell, of the syndrome to which he gave the name "causalgia," meaning "burning pain," which is the characteristic feature of the condition. The problem arose acutely in the Boer War and again in the first World War. The syndrome in its classical form is fortunately rare in civil life, but closely allied conditions have been described by Leriche under the heading "spreading neuralgia," by Homans as "minor causalgia," and by Livingston under "post-traumatic pain syndrome," and examples are relatively common.

The typical syndrome appears particularly as a sequel of gunshot wounds, often of an apparently trivial nature, where some superficial nerve or blood-vessel is involved. Local inflammation with subsequent scarring has frequently preceded the onset of pain. The pain has a burning quality and is persistent, with superimposed paroxysmal exacerbations, during which times it becomes unbearable, and may lead to suicide. The pain, at first, may be distributed within the territory of a particular nerve, but later spreads beyond this confine, and its distribution follows no known anatomical distribution of somatic afferent fibers. It is associated with extreme hyperaesthesia, sometimes located to special "trigger points." With the persistence of pain, there appears vasomotor, sudomotor and trophic changes. The skin, over an increasing area, becomes dry and glossy, bright red in appearance, with sometimes a rise of several degrees in surface temperature. In later stages, finger tips become tapered and atrophic, and intractable

ulcerations may appear. The nails become brittle, the finger joints stiff and swollen, and the limb is held rigidly by reflex spasm in skeletal muscles.

If untreated, the syndrome persists and tends to develop momentum, leading to intense suffering and finally to a complete breakdown of the psychic stability of the patient. The progressive nature of the condition can be realized best by the vivid clinical picture given by Weir Mitchell:

"The great mass of sufferers described this pain as superficial, but others said it was also in the joints and deep in the palm. If it lasted long, it was finally referred to the skin alone.

"Its intensity varies from the most trivial burning to a state of torture, which can hardly be credited, but which reacts on the whole economy, until the general health is seriously affected. The part itself is not alone subject to an intense burning sensation, but becomes exquisitely hyperaesthetic, so that a touch or a tap of the finger increases the pain. Exposure to the air is avoided by the patient with a care which seems absurd, and most of the bad cases keep the hand constantly wet, finding relief in the moisture rather than in the coolness of the application. Two of these sufferers carried a bottle of water and a sponge, and never permitted the part to become dry for a moment. As the pain increases the general sympathy becomes more marked. The temper changes and grows irritable, the face becomes anxious, and has a look of weariness and suffering. The sleep is restless, and the constitutional condition, reacting on the wounded limb, exasperates the hyperaesthetic state, so that the rattling of a newspaper, a breath of air, the step of another across the ward, the vibrations caused by a military band, or the shock of the feet in walking, give rise to increase of pain. At last, the patient grows hysterical, if we may use the only term which describes the facts. He walks carefully, carries the limb with the sound hand, is tremulous, nervous, and has all kinds of expedients for lessening his pain. In two cases, at least, the skin of the entire body became hyperaesthetic when dry, and the men found some ease from pouring water into their boots. They said when questioned, that it made walking hurt less; but how, or why, unless by diminishing vibrations, we cannot explain. One of these men went so far as to wet the sound hand when he was obliged to touch the other, and insisted that the observer should also wet his hand before touching him, complaining that dry touch always exasperated his pain."

There can be little question that the psychological make-up of the

individual and specific situations of conflict must play important roles in the pathogenesis of causalgia. A full psychiatric study of such individuals has curiously never been undertaken. The patients are said to possess the kind of "temperament" that predisposes toward psychoneurosis, yet it has been repeatedly emphasized that, before injury, they have usually shown no emotional instability whatever. Leriche describes one soldier as "gay and full of courage" when he arrived in hospital, who subsequently became "the most unbearable individual by reason of his intense suffering." The immediate return to psychic stability when the pain is relieved, even temporarily, has also been commented upon.

Perhaps the most striking feature of the story is the dramatic and lasting relief of pain which, at least in minor causalgia, may follow large infiltrations with novocaine locally, particularly of the "trigger points." This observation has been made independently in several clinics throughout the world. The psychic implications of such a method of treatment cannot be lost sight of, but it is a fact that many of these patients have had no idea that the infiltration would give them any relief, and have submitted to it as a matter of routine examination. The relief of pain has come as a surprise to them. Following the novocaine infiltration, the vasomotor disturbances and reflex spasm of skeletal muscle may disappear almost instantaneously. These are changes which can be recorded objectively.

The role in etiology of a focal point of irritation at the periphery is obvious. Livingston* has elaborated the view that such a focus sets up a constant bombardment of spinal centers, resulting in reflex vasomotor and skeletal muscle changes, which, in turn, may reactivate the peripheral irritative focus, and thereby set up a vicious circle. The consequent interneuronal activity within delayed and possibly reverberating (self-reëxciting) paths establishes a rhythm which is maintained by excitation, both from the periphery and from "higher centers," giving rise by summation to lowered synaptic resistance and thereby to ever-widening reflexes and to continual spread of the pain.

This is only one rather vivid example of the psychosomatic factor which plays a role in all disease processes. Functional disturbances can lead to irreversible structural change. Such functional disturbances are frequently only the exaggerations of the "normal" fluctuations occurring in physiological mechanisms. It behooves us therefore to pay closer attention to these "normal" fluctuations and to study

* Pain mechanisms. Macmillan. New York. 1943.

them under varying degrees of emotional stress Today this is a matter of national urgency.

"The question of the relation between the working of the brain and the working of the mind," writes Sherrington, "is, we hear often, one improper to put" But he continues, "only *after* the question has been discussed, can they (the psychiatrist and the physiologist) go on their respective ways, as perforce they ultimately must, disappointed it may be, but wiser, if sadder, practitioners and men"

DISCUSSION OF THE PAPER

Prof. C. W. Hampel (*College of Medicine, New York University, New York, N. Y.*):

Dr Sheehan has given us a clear picture of the arrangement of the functional units of the nervous system for dispersal of nervous influences, and has pointed out that this arrangement permits a variability in the response The extent to which the response may vary under apparently identical experimental conditions is not always fully realized

Appreciation of this arrangement helps us to understand why the opposing forces which operate to attain or maintain an equilibrium do not always act with equal directness An example of this principle is to be found in the inhibition of the crossed extensor reflex The extensor muscles, thrown into contraction by contralateral stimulation, are made to relax immediately by the inhibiting influence of ipsolateral stimulation in spite of the continuance of the excitatory contralateral stimulation Relaxation can of course be brought about also by cessation of the contralateral stimulation The effect, however, is not immediate but is characterized by an after-discharge The maintenance of postural contractions may be looked upon as the resultant of opposing forces which act with unequal directness on the motor units

In responses involving the autonomic nervous system, the same sort of phenomenon may be observed As the blood sugar level of a cat is slowly lowered through the action of insulin, a point is reached at which the sympatho-adrenal mechanism is called into play The diffuse nature of the sympathetic discharge is clearly evident in the pupillary response, heart rate, and other familiar signs The effect of the adrenin on the blood sugar is a rapid one, and, if the glycogen stores of the body are ample, may in a few minutes restore the blood sugar to a level that will require the continued action of insulin for an hour or more to depress it to the critical level again

The influence of previous experience upon the subsequent responses of the organism may be seen in studies of the mechanisms that operate to maintain a steady state in the internal environment Examples attesting the importance of training on the efficiency of homeostatic mechanisms are numerous The convalescent, impatient to be about after a long illness, finds that his vasomotor system requires a period of re-education before he can assume the erect posture or maintain it for any length of time It is apparent, moreover, that this training need not necessarily be specific Schneider has found that a man's "ceiling"—his ability to withstand lowered oxygen tension—was definitely related to the state of physical fitness as measured by the efficiency of his cardiovascular responses to exercise In this instance, a high degree of physical fitness maintained by controlled activity and rest on the ground could prepare a man for activity at high altitudes The effect of a person's previous experience or training on his ability to perform in another field of endeavor is not a new observation, but we are just beginning to appreciate the basic physiological principles involved in the responses of homeostatic mechanisms to extreme changes in the environment

Dr. Bela Mittelmann (*Cornell University Medical College, New York, N. Y.*):

It is customary to group psychosomatic disorders according to the respective organs that are seriously disturbed in their function. Thus it is customary to speak of gastrointestinal disorders such as ulcer or colitis, of asthma, hypertension, urticaria and migraine. The recording of various physiological functions in patients suffering from these various disorders, however, shows that besides the mainly affected organ, other functions change also during emotional stress. Thus in patients with peptic ulcer the finger temperature and the pulse rate change during conflict, anger and anxiety. For this reason it is more correct to say that physiologically patients react with an extensive range of physiological functions during emotional stress; of this extensive range in various patients, one or another is predominantly disturbed during stress. This formulation is in harmony with Dr. Sheehan's emphasis on broad approach to the neurophysiological patterns, in place of thinking of them in unicellular terms.

What I have said about the physiological aspect applies to the psychopathological side also. It is customary to say that certain psychosomatic disorders occur in individuals with a specific type of personality structure. Thus, migraine occurs in perfectionistic individuals, hypertension in individuals with strong hostility, asthma in dependent individuals. Similarly, we found in our investigation that peptic ulcer occurred mainly in perfectionistic, hard driving, ambitious individuals. I would like to accent the word *mainly*. We have found peptic ulcer to occur in dependent individuals also and in individuals who in the conduct of their lives were the reverse of ambition and perfectionism, one of our subjects having been a professional beggar. A consideration of the special conflict situations during which the symptoms become most acute gives similar results. Ulcer patients most commonly broke down with symptoms if they failed in their ideals of perfectionism and independence, and thus their anger, anxiety and guilt were aroused. This too, however, was not universal, and we have found individuals whose symptoms occurred when they were frustrated in their dependency longings to which they openly and defiantly confessed. Thus we can say that the individual reacts with a variety of personality and emotional patterns; to given situations of this variety, certain personality features and conflict patterns predominate in characteristic psychosomatic disorders.

This is not surprising if we realize that certain types of attitudes are present in nearly every normal and pathological psychological reaction. A measure of needing support from other human beings, a measure of ambition, need for self-esteem and anxiety in certain situations is present in every person and fear of rejection, hostility, guilt, disturbance of significant organ functions is present in every psychopathological state from schizophrenia to anxiety hysteria. This does not mean that the mentioned disturbances are identical but they invariably have *common* features as well as predominantly *specific* features, with considerable latitude and variety in patterning.

These principles were dramatically illustrated by a patient with Raynaud's syndrome on whom we conducted extensive experiments. She had characteristic attacks of pain and cyanosis of the fingers. These attacks started while she was pregnant, when her second husband began to mistreat her as the first one did. She became hostile, anxious and sexually frigid. Thus she presented a personality and a conflict pattern characteristic of the symptom of cold extremities; a dependent individual with conflict over hostile and sexual impulses with anxiety and guilt. This patient underwent a bilateral cervicodorsal sympathectomy which relieved her symptom However, a year later she returned to the hospital suffering from peptic ulcer.

I may add that it is of considerable practical significance to realize that in all disorders we are dealing both with general and specific problems of pathophysiology and psychopathology. In the New York Hospital we have been engaged in recent months in the attempt to determine the presence or susceptibility to gastrointestinal disturbances during emotional stress in military life. We first started by concentrating on the personality features, conflict patterns and the symptoms considered specific for these disorders. Soon, however, we found that we missed too many individuals. Our results became much more reliable when

we developed methods to essay most of the individuals' significant reactions as regards psychopathology and pathophysiology in general, with special emphasis on whatever psychosomatic disorder we were interested in I may summarize by saying that in all psychosomatic disturbances there are general and special problems of psychopathology and pathophysiology with a considerable latitude of patterning The best results in quick selection of patients is obtained if a survey of all significant psychological and physiological disturbances is combined with that of the specific features

DISTURBANCES OF GASTROINTESTINAL FUNCTION IN RELATION TO PERSONALITY DISORDERS*

By Harold G. Wolff

Cornell University Medical College, New York, N. Y.

Previous studies on normal subjects and patients with peptic ulcers have shown that day-to-day life situations that provoked certain patterns of emotional reaction induced hypersecretion in the stomach comparable to that resulting from prolonged absorption of histamine, vagus stimulation and sham feeding. During periods of experimentally induced anxiety, hostility and resentment, we have found a rise in acidity and increased contractions in the stomachs of all the patients suffering from ulcer and in many normal subjects Moreover, it has been possible to reverse this process and cause a decrease in acidity and motility by inducing in these patients feelings of contentment and well-being In these patients, a history of prolonged emotional turmoil involving mainly conflict, anxiety, guilt, hostility and resentment has been found

A man, aged 57, who has fed himself since the age of 9 through a surgically produced permanent gastric fistula, is the subject of this report It is his custom to put food into his mouth and, after tasting and chewing it, to expectorate it into an ordinary kitchen funnel inserted into his stoma.

Various measures of gastric function were used Color changes in the gastric mucosa were compared to a standard color scale. Output of acid by the parietal cells was estimated Records of stomach contractions were sometimes made. Estimates of the subject's emotional state were recorded, not only during the experiments but also at separate daily interviews, and these were classified according to the dominant elements. The emotional reactions were then correlated with the changes in gastric function.

While acid was continuously elaborated in the subject under basal conditions, and spontaneous transitory acceleration occurred, undue and prolonged acceleration of acid secretion in the stomach, however provoked, resulted in hyperemia and engorgement of the mucous membrane resembling hypertrophic gastritis.

*The material of this paper has been published as follows: Emotions and gastroduodenal function, Mittelmann, Bela, & Wolff, Harold G., Psychosomatic Medicine, 4, 1942; Evidence on the genesis of peptic ulcer in man, Wolf, Stewart, & Wolff, H. G., Jour Am. Med. Assoc., 120, 1942. Human gastric function, Wolf, Stewart, & Wolff, H. G., Oxford University Press, 1943.

In the spontaneous transitory phases of accelerated secretion, blushing of the mucus membrane and vigorous contraction of the stomach wall were observed, but such emotions as fear and sadness were accompanied by pallor of the mucosa, and inhibition of contractions and acid secretion.

Accelerated acid secretion, hypermotility, hyperemia and engorgement of the gastric mucosa accompanied emotions of anxiety, hostility and resentment.

Sustained anxiety, hostility and resentment were accompanied by severe and prolonged engorgement, hypermotility and hypersecretion In this state, the mucosa were unusually susceptible to injury, and the pain threshold was lowered, so that otherwise painless contractions, acid concentrations and mechanical stimuli became painful.

The mucosa were ordinarily protected from injury by an effective coating of mucus, but when this was lost even minor traumata caused oedema, inflammatory changes, erosions and hemorrhages.

Further hyperaemia and acceleration of acid secretion was induced by contact of acid gastric juice with a denuded surface, and such prolonged contact resulted in the formation of a peptic ulcer.

Apparently the natural history of peptic ulcer in human beings involves a chain of events beginning with anxiety and conflict and the associated overactivity of the stomach and ending with hemorrhage or perforation.

THE RORSCHACH METHOD IN THE STUDY OF PERSONALITY

By M. R. Harrower-Erickson
University of Wisconsin, Madison, Wisconsin

Although I have spent perhaps the best part of my waking life during the past five years mulling over some problem connected with the Rorschach method, I still remember keenly my first reaction to the test It was one of incredulity and disbelief. Just as Naaman, the king in the Old Testament story, refused to believe that anything as simple as bathing in the River of Jordan could cure him of his disease, so I refused to believe that the task I had been set, to look at ink blots, could provide the examiner with far-reaching information about personality in general, and my own in particular! And when this examiner, a friend of many years, looking over my responses, remarked: "Ah! How little I knew you"—this seemed to add insult to injury.

Any of you who are hearing about the Rorschach method for the first time are entitled, therefore, to just such a reaction of incredulity! On the other hand, those of you who are well informed as to the intricacies that lie behind the simple facade of the test, who are experts in your own right, must bear with me while I make the necessary introductions.

In 1921, Hermann Rorschach,[1] a Swiss psychiatrist pointed out that if you showed people a series of ink-blot pictures (like the 10 he had devised after many years experimenting) you would find that they all saw in these meaningless blotches a multitude of diverse objects—the blots, or parts of the blots would "look like" things to them Your learned academic friends, the old janitor, the patients in the State Hospital, all would have meaningful experiences when looking at these calculatedly meaningless blotches. Moreover, Rorschach explained, this meaning, mental organization or sense which was given by the individual to the non-sense, was never the result of chance but was directly related to his ways of acting, his patterns of behavior, his personality.

The Rorschach test, therefore, consists in the subject describing to the examiner what he sees in the blots and the examiner writing down these responses, the finished product being the Rorschach record.

Unlike previous investigators who have worked with ink blots, and

[1] **Rorschach, H.** Psychodiagnostik. 1st ed. Ernst Bircher, Bern. 1921.

who have stressed the importance of the association in the *content* of the responses, or who were concerned with the *what* was seen, Rorschach focused attention on two other aspects: on the *where* in the blot the perceptual object was seen, and, secondly, on what *perceptual quality had been decisive in evoking* the response; that is, was the individual swayed, let us say, by the similarity of form, color, or apparent texture of the blot to the object that he had in mind.

Let us make this important but somewhat abstract distinction more concrete, and at the same time give you the first-hand experience of "seeing things" in the blots. In FIGURE 1 we have an uncolored reproduction of card VIII in Rorschach's ink-blot series. As you look at it what does it suggest to your mind?

Rorschach would be concerned not only with finding out *what* the blot appears like to you, but *where* your experiences lie. Do they involve the whole blot? Could it be a coat of arms? Do the side areas suggest an animal perhaps? Is this an animal because it has the shape, say, of a bear or beaver, or because of the incipient feeling of movement in the outstretched front paws, or because it has some kind of furry texture that can be seen, and so on. Or does the area in the center resemble two flags?

Impressed by these varied facets of any one answer, and in order to make an analysis of each response objective and accurate, Rorschach developed a method of describing or scoring the ingredients of each perceptual experience. This, while it sometimes appears to the outsider as unnecessarily technical and esoteric, has actually been the step which has made objective and uniform studies possible. When we speak of a Rorschach record *which has been analyzed*, therefore, we mean that the listed responses have been translated into the symbols that describe their perceptual components, which, in turn (when the various totals, ratios, and calculations have been determined), enable us to gauge the presence or absence, strength or weakness of the more general psychological ingredients that, interestingly and strangely enough, are their counterparts.

What are some of these more general traits or ingredients of personality? They include the individual's self-control, the rapport or positive emotional relationship that he has achieved with his fellow men, his inner poise or stability, his anxieties, the strength of his more primitive drives and emotions, the integration or lack of integration of the personality as a whole.

One way to describe the type of information which the test gives us is to liken it to a mental X-ray picture in that the psychological skele-

Figure 1

Uncolored reproduction of card VIII in Rorschach's ink-blot series

ton is laid bare and concealed weaknesses or broken psychological bones are demonstrated. I prefer a somewhat more dynamic analogy, however, as, for example, that we are looking at the dials or indicators of some big power house. On such dials are recorded the relative strengths of various drives, needs, impulses, instincts—call them what you will—and also the resistance of the various controls or disciplines that keep these energies in check. The sum total of these controlling mechanisms and energies may result, in some cases, in smooth-running function; in others the pressure may burst the controls and cause an explosion; or again, there may be no pressure, both in power houses and in individuals.

With this last analogy in mind, let me now introduce to you the so-called psychogram (FIGURE 2), a graphic device for epitomizing the scored Rorschach record, first introduced in the *Rorschach Research Exchange* and slightly modified in the figure to include the maximum information. Here, the individual's total number of responses are charted in terms of the scoring symbols of Rorschach of which I spoke, with additional components introduced by B. Klopfer, which are now employed by a large number of investigators.[2] That is, the responses are charted in terms of where they are seen in the blot, and in terms of the qualities which evoked them (see legend). It is also a good illustration of the power-house analogy, for here are the various psychological ingredients demonstrating their relative importance in the total psychological make-up.

Hence, we now possess a tool that enables us to gauge an individual's psychological dynamics. Now the question is: Where can such a tool be used?

From the wide choice of fields where the test is now employed, I have selected two for discussion: its use as an aid in clinical diagnosis on the one hand, and examples of its use in large-scale investigations of various kinds where some type of screening or selecting is aimed at on the other hand, these large-scale investigations having become possible in the last year or so through the development of the group Rorschach.

The use of the test clinically of necessity requires the assumption that the psychic and somatic conditions of the patient are interdependent. To make a diagnosis or a prognosis of the patient's condition on the basis of psychological characteristics would be nonsensical if the relation of psyche and soma were an arbitrary one.

[2] **Klopfer, B., & Kelley, D. M.** "The Rorschach Technique." The World Book Co. Yonkers. 1942.

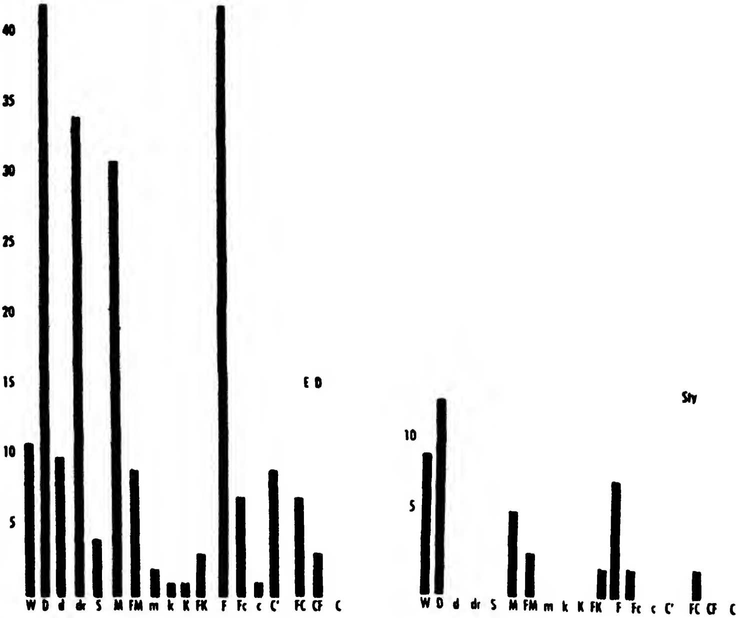


FIGURE 2. This figure represents two psychograms of "normal" individuals. E. D. is the record of a superior normal, i.e., an individual with a college education and a successful artistic career. Stv. is the record of a student nurse just out of high school. While both records show good adjustment in the distribution of the responses, that of E. D., if we consider only quantity, shows much greater productivity.

The columns denoted by W, D, d, dr and S refer to the places in the blots in which the responses were located. For example E. D. has 11 responses in the W column; this means that 11 of the responses involved the use of the whole blot; 42 responses were in the D areas (large details), 4 in the white spaces between the blots, and so on. In the columns M through to C' we have the same responses again recorded, but this time in terms of the qualities which determined their selection. For example, the 7 FC' responses in E. D.'s record show that seven choices were based on a combination of the form of the blot and its color. Full discussions of the scoring will be found in references 1 and 2.

One of the most striking examples of a specific Rorschach personality pattern that is correlated with specific physical abnormality is the case of gross cerebral pathology, as tumor, widespread atrophy and scarring. In these cases, there is almost invariably a personality change clearly demonstrable by the Rorschach, even in those cases where neurological examination may be essentially negative, and where intelligence, as measured by the regular intelligence tests, is superior.

You will remember the appearance of the normal individual's psychogram, as seen in FIGURE 2. This should be compared with the im-

poverished personality of the patient with some cerebral lesion as seen in FIGURES 3 and 4.

I have had the opportunity to obtain more than 100 such records from patients in which a cerebral lesion has been demonstrated at operation. In many cases, of course, no diagnostic problem existed, the clinical evidence and encephalogram being conclusive, but in others, surprising as it may seem, this personality change has been found where organic pathology may not have been suspected and where a real problem in differential diagnosis existed. In a recent attempt to introduce localizing evidence, it has seemed to Dr. Theodore Erickson that just that particular constellation of no neurological signs, together with what might be described as a very "positive" Rorschach, was indicative of a frontal lobe lesion.

Many interesting cases could be cited. An individual in one of the services, referred for psychological examination because of complaints that had seemed to be of a hysterical nature, showed a psychogram of the "organic" type The electroencephalogram demonstrated a large

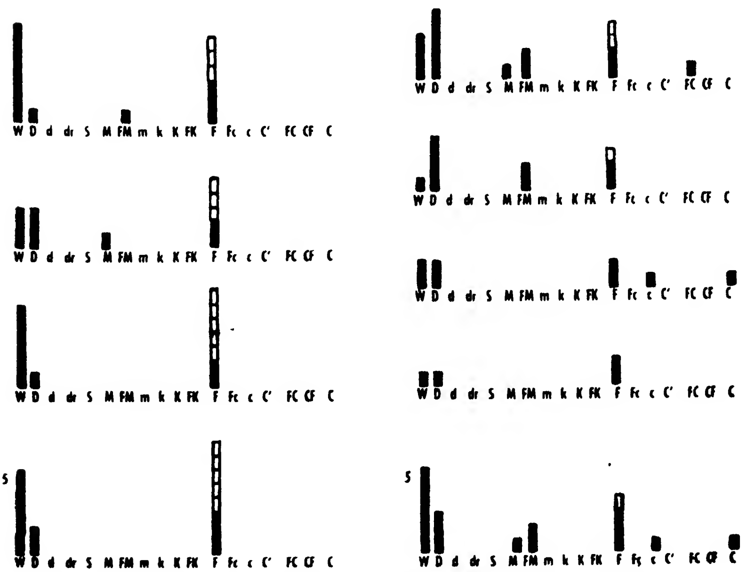


FIGURE 3 Psychograms of 9 patients with cerebral lesions. The extremely small output (no record has more than 8 responses) and the uniformity of the records should be noticed Other characteristics, not found in normal records, are the so-called F minus answers; that is, answers not in any way justified by the actual form of the blot. These are represented in the psychogram by the white segments in the F column (See reference 3 for further details.)

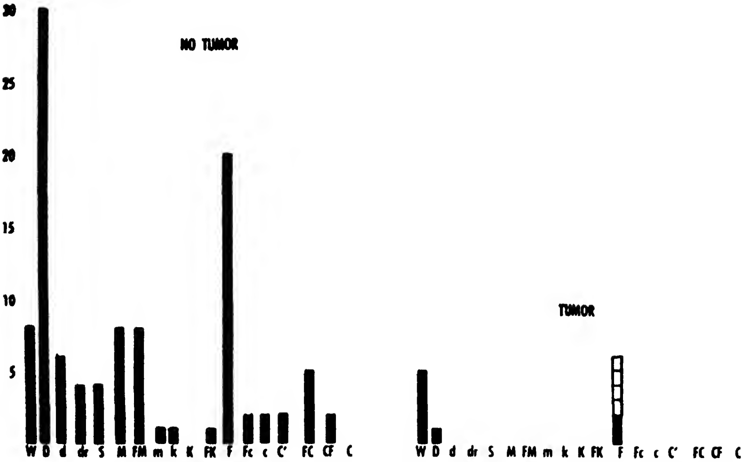


FIGURE 4 Cases for differential diagnosis Records of two patients referred for Rorschach examination with the provisional diagnosis of frontal lobe tumor The record on the right can be seen to be of the type demonstrated in FIGURE 3 This patient was found, at operation, to have a frontal lobe lesion The other record shows a normal personality, and the final clinical diagnosis in this case was not that of cerebral tumor

area of abnormal waves, and a history of cerebral trauma, previously considered unimportant, was brought to light

An interesting comparison sometimes occurs when two patients markedly euphoric are, at first, *both* considered clinically as suspects for frontal lobe tumor. The Rorschach record of the one, however, may show that the euphoric symptoms are related to the impoverished personality of the patient with some intercranial pathology[3] and may be seen as the vacuous attempts of the individual, handicapped by reduced psychological equipment, to retain his place in the environment. The other record may show in the profusion of Rorschach responses, and its totally different psychogram, the runaway energies of the manic phase of psychosis. Such diagnoses will be confirmed at operation, in the one case, and by continued negative neurological findings and negative encephalogram, in the second

The *converse* of this type of diagnosis may be found, for example, in the patient with complete hemiplegia whose Rorschach record, however, does not indicate cerebral trauma, but does indicate an acute psychological disturbance. This takes us to another important con-

[3] Harrower-Erickson, M. R. Personality changes accompanying cerebral lesions. 1. Rorschach studies of patients with cerebral tumors. Arch. Neurol. & Psychiat. 43: 859–890. 1940.

tribution of the Rorschach, the question of detecting cases where psychogenic factors are involved.

One of the most frequent questions put to the Rorschach examiner by physicians in my experience is whether, from the personality pattern elicited, it would seem that psychological factors are contributing to, or causing physical symptoms. The patient may be hemiplegic, may have convulsions, ulcer symptoms, hypertension, headache and vomiting, backache, but somewhere the suspicion has been aroused that such symptoms are not the whole story. The contribution of the Rorschach can be quite important. It is a *quick* way to direct the line of future investigation by the physician. If, for example, the personality is profoundly inadequate, maladjusted, unbalanced, the chances are high that this maladjustment plays some part in the total psychosomatic picture If, on the other hand, the patient demonstrates a well-balanced, rich, stable personality, such a finding would be an indication that the root of the trouble may still be unearthed by physical examination. These findings can never be 100 per cent certain, for, as W. D. Ross[4] has pointed out, the sociological or environmental factor must always be considered. For example, under undue stress from environmental factors, the essentially stable personality may show signs of psychoneurosis clinically.

Again, I must be content with a few specific examples FIGURE 5 shows the psychograms of two patients with very similar symptoms—pain in the arm. In the patient whose record is shown in FIGURE 5a, the Rorschach is evidence against an impoverished or maladjusted personality being an important factor, and, in this case, finally a small neuroma was found and removed at operation, leaving the patient quite free from pain. In the case of the patient whose record is shown in FIGURE 5b, the pain was found to be largely of psychogenic origin, involving the disinclination to return to a complicated home situation.

Or, one could quote as an example another patient (a former nurse who had cared for epileptic patients) with exquisitely focal seizures of four years' duration. On the Rorschach, she demonstrated a profound psychological disturbance indicative of sexual trauma. Psychiatric examination revealed just such a traumatic experience: rape, shortly before the onset of the seizures. The convulsions ceased completely following psychotherapy, and she has had none for over three years.

I have recently compiled the Rorschach findings of 100 consecutive

[4] **Ross, W. D.** The contribution of the Rorschach method to clinical diagnosis. Jour. Ment. Sci. July, 1941.

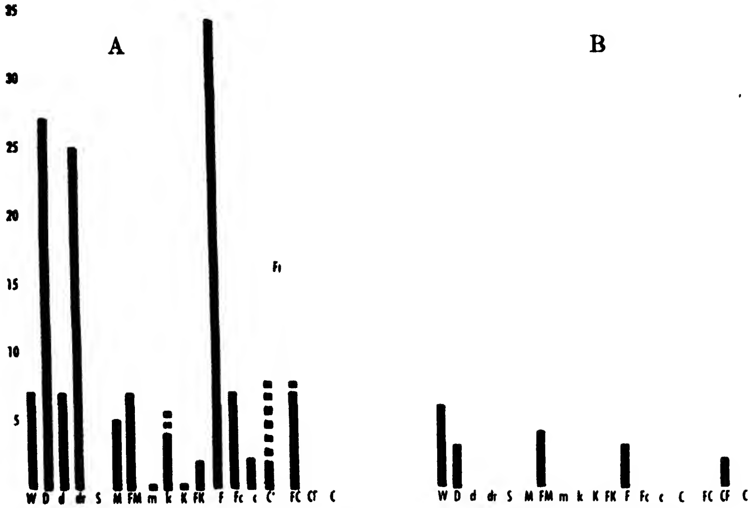


FIGURE 5. Cases for differential diagnosis A, no psychogenic factors; B, hysteria.

cases referred with this question of whether or not psychogenic factors can be said to be playing a part.[5] Certain Rorschach signs of maladjustment (which Mrs. F. R. Miale and I previously isolated in the records of patients considered as neurotic[6]) have been looked for in the records of these 100 patients and in the records of 345 control subjects, i.e., unselected healthy individuals in various walks of life. When five or more of these nine signs or characteristics can be shown to be present in a record, we have a very strong suspicion that psychogenic factors play a part. The "sign" approach is analogous to that introduced by Piotrowski as one way of evaluating the organic record.

Another way of bringing out these results can be seen in FIGURE 6.

A group of patients who have been of special interest recently have been those with ulcer symptoms. Both Dr. Ross and I had the opportunity at the Montreal Neurological Institute last year to examine soldiers invalided home from England with ulcer symptoms. We both found that while no specific personality pattern could be said to be correlated with the symptoms, that incidence of the so-called neurotic signs was high in these records. Moreover, we found that these signs

[5] Harrower-Erickson, M. R. Diagnosis of psychogenic factors in disease by means of the Rorschach method. Psychiatric Quarterly 17: 57–66. 1943.

[6] Miale, F. R., & Harrower-Erickson, M. R. Personality structure in the psychoneuroses. Rorschach Research Exchange 4: 71–74. 1940

TABLE 1

No.	Classification	Per cent with 5 or more "neurotic signs"
54	Neurotics	85
20	Neurotics with somatic disturbances .	65
26	Somatic disturbances only	15
108	College students, male	5
46	College students, female	11
40	Nurses in training	5
44	Aviation cadets	16
20	Superior adults	0
20	Superior adults	10
41	Convicts (I.Q range 77–141)	27
40	Orderlies in R.C.A M.C. (C grade)....	38
385	Controls	15
74	Neurotics	80

appeared equally frequently in patients who were finally diagnosed and discharged as true ulcer cases, that is, in whom X-ray had revealed evidence of the ulcer, as in those who were discharged as being

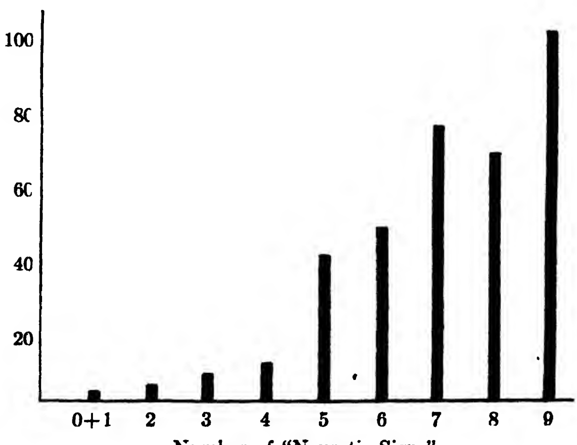


FIGURE 6 Percentage of neurotic cases found among records showing the various numbers of signs Total number of records 459
It should be noted that while only 2 per cent of those in the group with o or 1 sign were neurotic, 100 per cent of those found in the group with 9 signs were neurotic patients. Another interesting feature is the break between 4 and 5 signs, 39 per cent of those with 5 signs being diagnosed as neurotic clinically while only 10 per cent of those with 4 signs were so diagnosed.

"psychological" cases, i.e., those with no X-ray evidence. In view of the extremely interesting work recently reported by Dr. Wolff and Dr. Mittelmann, these findings perhaps have added significance.

Anyone who has worked with the Rorschach clinically cannot fail to be impressed by its possibilities. It was an obvious step, therefore, in 1939, when Canada entered the war, for those of us at the Montreal Neurological Institute to direct our energies toward utilizing it in some way in the war effort.

One of the obvious disadvantages to the widespread use of the method was the fact that, as it stood, it was too time-consuming a procedure for administration to large numbers of persons either for screening or selection. With the help of Miss M. Steiner, I therefore spent some time in an attempt to obviate this main difficulty. After systematic experimentation we developed the so-called *group Rorschach* which, utilizing slides of the original cards and requiring the writing of the responses in specially prepared booklets, allowed several hundred persons to be examined simultaneously in one hour.

It is not relevant now to go into the various ways in which we satisfied ourselves and, finally, I think, the most skeptical Rorschach workers, that what we derived from this method was, except for minor nuances, essentially the same as from the individual test. The results that I shall discuss next, however, will have been derived from the group Rorschach investigations in various fields, including personnel work in colleges, screening in mental and penal institutions, and studies in vocational selection.

The relation of the student's personality to his success or failure in college is a challenging field. Dr. Ruth L. Munroe,[7] a pioneer worker, has for the past few years given a Rorschach test to every member of the entering freshman class in Sarah Lawrence College, for the last two years, using the group method. Her analysis of each individual student has been available in the event that detailed information was needed, but she has also grouped her students into several categories on the basis of their performance on the test, so that these can be compared, on the one hand, with the academic results at the end of the year, and, on the other hand, with the evidence of acute behavior problems needing psychiatric advice.

In the past two years, with the collaboration of Dr. W. D. Ross and Mrs. Helga Malloy, I have had the opportunity to do the same thing with all entering freshmen in the medical school at McGill University.

[7] Munroe, R. L. An experiment in large-scale use of the Rorschach method. Jour Psychology 13: 229–263. 1942

This is a project planned for four years, which also includes EEG studies by Dr Jasper The results are now available on midyear and final examinations of the first year I think that TABLES 2, 3 and 4 will be a good basis for showing the progress of the work to date

Another interesting experience with the group technique has been the examination of psychotic patients in several mental hospitals Contrary to expectation, I have found, with one or two unruly excep-

TABLE 2

Dr Munroe's Ratings of Sarah Lawrence Students

	A*	B	C	D
Personality evaluation	23	23	15	21
Saw a psychiatrist	0	0	1	6
Academic failure	1		6	5

* A and B refer approximately to good average personalities and C and D to personalities showing questionable or pronounced difficulties

TABLE 3

Medical Students McGill University*

Personality rating	Number	Work good or Work satisfactory	Poor work Several failures Dropped
Excellent Above average Average Just below average	94	86% (81)	14% (13)
Poor and very poor	14	7% (1)	93% (13)

* The manner in which these results were obtained may perhaps be mentioned The names of all those who did outstanding work and of those who did extremely poor work (failed in many subjects or were dropped) were sent to the Rorschach examiner by the secretary to the faculty of [illegible]

mentioned by name Among the 86 per cent therefore are included the names of those who did exceptionally good work and those whose names were not mentioned

TABLE 4

Prediction*	Results
Basic potentialities better than 'poor' group but at present *disturbed* and *anxious* If this continues likely to produce d e t r i m e n t a l effect on studies If they can work out of their difficulties they will probably make the grade'	Cu Failed all subjects "Worried and could not attend" Referred for psychiatric examination *anxiety state* diagnosed Ca Age 40 Has wife and family and serious financial worries "Carried outside employment Extensive lung shadows found over which he worried a great deal"

* Sometimes specific predictions were verified in an interesting way as for example those that commented on an individual s anxiety

tions, such patients are able to participate in group procedures. In fact, as one medical officer explained to me, it is apparently easier for some of the patients to take the test in this way because the rapport which they achieved with the group allays some of their suspicions and negativistic attitudes toward such an examination. From these investigations, we have derived the all-important statistical material for comparative norms of various kinds for the group procedure. It has also been possible to derive from the new technique *new diagnostic* features pertaining to certain psychopathological entities. These are perhaps a little technical to be discussed here, but they have already proved of value in screening out the unsuspected psychotic in the unselected group

A cross section of the population at a penal institution affords a good medium for demonstrating the screening potentialities of the Rorschach test. Last year, owing to the kind co-operation of Dr. Banay, psychiatrist at Sing Sing Prison, I had the opportunity of examining a group of prisoners that he had picked for various reasons, but concerning which I knew nothing. This investigation yielded some very interesting results which I may mention briefly:

1. It demonstrated that the group procedure is a perfectly suitable test for persons of below average intelligence, a question that had been raised when we initially presented it. Half the group examined at Sing Sing had I.Q.'s below 100 and over 25 per cent fell in the 70–90 range. In all cases adequate records were obtained.

2. It is probably hardly necessary to mention that no typical "criminal personality" came to light. The records of the 40 prisoners were as diverse and individual as any other group of a similar number might be, some demonstrating very profound psychological disturbances, a few with extremely well-adjusted personalities.

3. However, when analyzed statistically, certain abnormalities for the group as a whole stood out clearly, chief of these being the marked predominance of the explosive and more primitive type of emotional responses over the more adjusted and well-integrated ones.

4. Certain gross abnormalities could be spotted immediately; for example, an "organic" personality pattern in the case of an individual with lues of the central nervous system, an incipient schizophrenic pattern in the record of an individual recently released from a state hospital, where he had received insulin treatment. An acutely disturbed record (showing perhaps the most alarming picture of the whole group at that moment) transpired to belong to a youth who had made a suicidal attempt that morning and was considered clinically as pro-

foundly disturbed. A well-adjusted pattern, indicating, however, that the individual had passed through a period of acute strain and stress prior to the better adjustment, was found to correlate with a previously obsessed and agitated individual whose behavior had changed drastically after a successful lobotomy.

5. While no attempt naturally was made to predict the actual crime from the personality alone, yet certain personalities clearly "belonged" to certain types of crimes. The individual committed for assault, for example, gave evidence of what might be described as a primitive imbalance, as opposed to the highly complex personality structure, in some cases where the crime was complicated and premeditated

Recently Lindner and Chapman[8] have reported their results of a similar type of screening at the Lewisburg Penitentiary. They state that they have selected "from undifferentiated admissions, *without a single failure,* all those who were later found by the staff to require special attention." This would seem to me a little unduly optimistic and, since no figures are given, one does not know the numbers examined to date, but the technique is obviously of value in such gross screening experiments.

As an indication that the more favorable personalities may also be caught in this particular net, the study of Piotrowski[9] and Candee may be mentioned. These authors have emphasized the selectional possibilities of the group method, and were able, I believe, to make correct predictions *as to the particularly competent mechanical worker* in 88 per cent of the 78 cases examined.

It would be misleading to give the impression that the Rorschach test, either in its individual or group form, is making a widespread contribution to the problems of personnel selection in the armed forces at the present time It would be equally misleading, however, not to point to some of the situations where it is being advantageously employed.

I recently had an opportunity of talking at length with a member of the Australian Air Force, who was in close touch with the use of the group method in screening and selection experiments in the Australian Air Force. His reports were very encouraging both on the findings at that time and the further extension of their plans.

In this country, Dr. Barry Bigelow has tested naval aviators in Pen-

[8] **Lindner, R., & Chapman, K. W.** An eclectic group method Rorschach Research Exchange 6: 139–146. 1943.

[9] **Piotrowski, Z.** Use of the Rorschach in vocational selection Jour Consulting Psychology 7: 97–102. 1943.

sacola and Jacksonville, prior to their training period, and follow-up studies on their subsequent performance in flying are available for comparison with their performance in the test. The results of these experiments have recently been made available to a few persons for intensive study from the Rorschach angle, and should greatly increase our knowledge of those specific personality patterns that are suitable for this type of performance and those that may have difficulties.

A similar investigation with paratroopers is now in its initial stages in Canada, which again promises to give us the kind of information we have been needing for so long. Under the direction of Dr. W. D. Ross, a group Rorschach is given to 100 entering paratroopers just prior to training. It has been felt by some people that the phenomenon of "freezing on the jump," which is shown in a certain percentage of students in each class, may well have a psychological origin that will be demonstrable on the Rorschach record. Dr. Ross's procedure, therefore, is to give the test to these students, and then to wait till the end of the six-weeks training period, when the records of those who "froze" are studied carefully. This will be repeated with a number of entering classes until a sufficiently clear-cut personality pattern has emerged or the high incidence of the neurotic signs has been established as significantly different in the "freezers." Predictions of "freezers" on the basis of the Rorschach record will then be attempted.

Hertzman and Seitz[10] have already published results of tests with the group method concerning changes in personality occurring at high altitudes and are making further studies along these lines.

Under the direction of Dr. Klopfer, group tests are now being given every two weeks to approximately 50 persons in the officer's training division of the U. S. Signal Corps in Philadelphia. A comment on this, recently received by Dr. Klopfer from the commanding officer, may be quoted:

"The use of the group Rorschach psychodiagnostic technique in evaluation of the qualifications of student officers for assignment is proving extremely valuable. When used in conjunction with our other psychological tests, it provides an opportunity to observe the interplay between intelligence and personality and to estimate the emotional stability of the officers under stress and responsibility. This knowledge allows us to recommend assignment according to the best interest of the service."

There are at the present time quite a number of trained Rorschach

[10] Hertzman, M., & Seitz, C. Rorschach reactions at high altitude. Jour. Psychol. 14: 245–257. 1942.

workers in the various branches of the services. These men have sometimes been able to make use of their training, while others have interested their superior officers in the possibilities of trying out the test in relevant local situations. From a number of camps and military hospitals throughout the country, requests for information on the test, and for materials for its administration, are constantly coming in.

Despite this obvious interest, however, it must, I think, be said in conclusion, that the vast resources of the Rorschach test have not yet been tapped. It is still the business of those of us who are convinced of its merits to inform those who are not yet acquainted with it of its potential value. Moreover, we must be sufficiently courageous to discard, if necessary, certain apparently sacrosanct features of the test if, by so doing, it can be reduced to a form that can be readily used where its need is vital.

DISCUSSION OF THE PAPER

Dr. Ruth L. Munroe (*Sarah Lawrence College, Bronxville, N. Y.*):

Large-scale application of the Rorschach test, using the group method of administration developed by Dr. Harrower-Erickson and some means of rapid assessment, is still a new venture. I think that I can contribute most to the discussion of this problem by reporting recent developments in our experimentation with such use of the test at Sarah Lawrence College. Perhaps the most immediate interest of our results is further confirmation of the validity of the group Rorschach I would like, however, to draw your attention to the special way in which we used the group test, because the results seem to have important implications for more general problems of selection and of test construction.

The Rorschach is a versatile instrument. The raw responses to the ink blots reflect so much of the personality that many lines of differentiation in personality analysis can be developed Until recently, the fully trained expert has used all the data in the test with as much clinical insight as he could muster. This is undoubtedly the ideal procedure, because the examiner has at his disposal fairly good norms for single items and key relationships of data. He also has a knowledge of how clinical syndromes are expressed in the test, which is unfortunately no more uniform from one subject to the next than the symptomatology of two schizophrenics is identical. Bringing all this material together into a sound diagnosis takes time and skill—the same order of skill that a psychiatrist must have in sorting the data of case history and interview into a significant picture of the personality.

Large-scale testing does not permit the necessary time for this type of evaluation, and large-scale testing must ultimately be done by psychologists less elaborately trained Moreover, the Rorschach is being used in new fields for new purposes. Efforts are being made in several directions, therefore, to reach more objective criteria for special conditions or special aptitudes. The "neurotic signs" developed by Dr. Harrower-Erickson represent one example of this trend. She found that diagnosed psychoneurotics actually deviate from normals more frequently on 9 items (technical Rorschach items like F and color shock) than on any others. Most neurotics and few normals have more than 4 of these "signs." Similar work has been done or is in progress on patients with organic brain conditions, on schizophrenics, on psychopaths, etc. Piotrowski and his colleagues have statistically isolated 3 "signs" important for success in shopwork. Dr. Bigelow and also Dr. Molich have compared the protocols of successful and unsuccess-

ful aviation cadets to determine objectively which items differentiate these groups most adequately. Such investigations can clarify, objectify, and, at times, for some purposes even supplant the general clinical evaluation described above. All of these efforts are directed, however, toward the diagnosis of *specific* conditions.

Before coming to the discussion of our different procedure at Sarah Lawrence, I should also mention the kind of work Dr. Harrower-Erickson has done in quickly predicting success in medical school. Indeed, I have done it myself in giving to students a rating that I called specific academic prediction. Dr. Harrower-Erickson obtained excellent results. Our results were also very good. Out of 45 ratings, aimed at predicting academic performance, 39 were "on the nose" according to the general average established for the student's work during the first year; only one was badly discrepant. Nevertheless, I feel that these ratings, however successful, did not adequately meet the necessary criteria for large-scale testing. They were highly composite affairs, based not only upon our knowledge of the students through their protocols, but also upon our awareness of the specific requirements of the academic situation. Our method was essentially "clinical," dependent upon our personal insight I am sure Dr. Harrower-Erickson would agree that future development of large-scale work should envisage both a more objective approach to the evaluation of the student and more precise knowledge of what the situation demands.

We have already tried to be more precise in one direction at Sarah Lawrence, though in a manner which may sound paradoxical and is certainly very different from the investigations mentioned above aimed at specific diagnosis. What we did was to give each student a quantitative rating on "general adjustment," *excluding* so far as possible her adaptation to specific academic requirements, indeed, excluding the selection of any particular type of personality (We will discuss, at some length, later the objective criteria in the test for this rating It seems preferable to describe first what it is and how it works.) Unintelligent and unintellectual girls, introverts and extraverts, aggressive and timid individuals, complicated and simple souls were all rated "adequate" provided the personality seemed to be functioning well. "Functioning well" meant initially—to be frank—nothing more than having a "good" Rorschach protocol. This criterion is, on reflection, pretty sound The test was developed by clinicians who knew mental disturbance in variety and had no particular axe to grind in defining normality Indeed, they did not define it at all, except "operationally" Reflection on the nature of the test suggests that what a good protocol means essentially is a reasonable balance or integration between the impulsive and controlling forces in the personality Control must be adequate but not excessive or too repressive. Great latitude is allowed in type and intensity of impulse and type of control, but their relationship must be sound In behavioral terms we defined adjustment very simply as the ability to "get along" reasonably well with reasonable inner comfort. Occasionally, we rated a girl badly adjusted with an asterisk to indicate that she gets along well, but at too high a cost to her own comfort

I shall depend primarily on our experimental results to show that this apparently vague concept of "general adjustment" does mean something that is empirically rather precise, useful, and measurable I must first describe the experiment. For two successive years (1940 and 41), we administered the Rorschach to the entire entering class at Sarah Lawrence College (225 girls in all), under carefully controlled conditions. Teachers were not informed of test results in order to guarantee complete independence of judgment. (Beginning this year the test is being used on a practical basis A feature of the work not presented here is a descriptive sketch of each student. Ratings and sketches are now available to teachers.) Evaluations from the test were made "blind," i.e., with no information about the student except her response to the ink blots. The test ratings were compared in June with the ordinary college records of academic performance and explicit notation of emotional difficulty—chiefly the list of girls brought to the attention of the college psychiatrist His advice is frequently sought by teachers in cases of minor maladjustment without referring the student directly

The adjustment ratings were very successful in predicting adjustment. Out of 100 girls rated "adequate" (A or B on a scale running from A to E), only 3 ap-

peared on any list of students in any sort of trouble, and 2 of these had minor upsets quickly solved. On the other hand, out of 33 students brought to the attention of the college psychiatrist, 30 had been rated as moderate or severe problems, 20 of them as severe problems. Many of these cases were not at all serious, of course, and with one striking exception, the Rorschach rating corresponded well with the psychiatrist's estimate of degree of difficulty. Ten of the 13 girls rated in the worst category by the Rorschach in 1941 either failed outright in their studies or had prolonged psychiatric attention, and the others were spontaneously described by teachers as rather neurotic.

Of greater interest to the present discussion, however, is the fact that 18 out of 19 students who were either dismissed or conditioned in their freshman year had poor adjustment ratings. Half of these girls were above the median on the ACE (American Council on Education Psychological Examinations, an intelligence test) one quarter above the 90th percentile Thus outright academic failure in the freshman year seems far more closely related to problems of adjustment than to lack of intelligence (This statement must not be reduced to the absurd. All entering students have a certain minimum of intelligence.)

The adjustment rating predicted degree of academic success *short of actual failure* as well as the intelligence test, but no better. Seventy-four per cent of the adjustment ratings and 71 per cent of the ACE scores (for purposes of comparability, the total distribution of ACE scores was reduced to 5 groups ranging from bad to good, numerically equivalent to the Rorschach ratings E to A) tallied with the academic average, excluding the cases of failure and conditioning The point that I find most significant, however, is the relationship between the two tests. Failures in the prediction of each measure can be at least partially explained by the other The small group of unadjusted girls who did satisfactory work all stood above the median on the ACE Conversely, with very few exceptions, the adjusted girls whose work was on the poor side stood in the bottom quartile on the ACE. *The two tests seem to measure demonstrably different things, both of which are important in academic performance.* An effort to *combine* them yields the following very suggestive results

1 When the two measures point in the same direction, good or bad, their combined predictive power is almost perfect No "adequately adjusted" student with a good ACE score failed No "poorly adjusted" student with a low ACE score did fully satisfactory work There were very few discrepancies with external measures of performance even of a minor degree.

2. Girls with "adequate" adjustment ratings and low ACE scores form a group which includes neither superior scholars nor outright failures Half of them proved to be weak students and several were rejected for return as juniors, although their work for the two lower years was considered passable Many of them made valuable contributions to the college as *people* and seemed to profit by their education as much as girls who got better grades In short, this group causes no serious trouble, but is likely to do mediocre work, at best, and, at worst, to trail along near the bottom academically

3. The most unpredictable group consists of girls with poor adjustment ratings and high ACE scores. This group contributes half of the dramatic failures and more than its quota of girls who just squeak by. It also accounts for several very superior students Statistically speaking, girls in this category are poor risks To eliminate them altogether is both impracticable, because there are too many, and undesirable, because one would eliminate the very good along with the very bad. These girls are probably the square pegs who need square holes, but an impressive number of them are well worth any special attention or tolerance required. Looking beyond the academic scene, it is probable that a good many distinguished, creative people—scholars, artists, aviators, etc, would do as badly on any general adjustment tests as the crackpots and dismal neurotics we would like to rule out.

Sensible procedure might be to avoid "speaking statistically" about this group and devote whatever time is available for selection to individual study of each case. Some types of maladjustment are likely to prove difficult in all situations More specific identification of what is wrong will suffice to cut out these cases.

Other types must be studied more carefully to determine whether their assets are especially important for the situation under consideration and their difficulties such as can be handled.

In spite of the fact that our evidence is too limited in scope to warrant safe generalization to other fields, it does, to my mind, suggest a useful hypothesis. Fragmentary observations strongly support the idea that the same problems obtain elsewhere. I should like to urge further experimentation with the concept of "general adjustment," as an empirical entity, to be measured separately and then combined with appropriate indicators of the special qualities required for any job in some such manner as that outlined above. It is not enough simply to screen out the mentally ill, and to measure special aptitudes or character traits independently. The concept of adjustment applied to the entire range of cases can probably be made to show, in a quick, practicable manner, the actual relationship between general personality factors and assets for a particular job. Prediction of success or failure could be made from a statistical combination of test scores with great accuracy in the majority of cases. The small group where errors are most likely to occur is isolated for more intensive study and the problem to be considered is clearly posed.

A further advantage of the concept is that the adjustment measure can be used in new contexts, as desired, with different sets of special data. A composite measure, oriented toward a particular situation but including personality factors (like our academic ratings and probably Piotrowski's signs), is less suitable for prediction in other fields. I should, perhaps, also emphasize the idea that "special data" could include not only aptitude tests, but also tests of personality configuration like the Rorschach itself differently analyzed, physiological measures and items from the case history.

That our results are not due exclusively to the magic of ink blots is shown by the fact that Mrs. Schnidl-Waehner obtained similar findings with an adjustment rating based upon her method of evaluating spontaneous drawings. Once the goal is clearly set, it should be possible to devise other techniques of measurement, possibly more practicable for large-scale use

This statement brings us back to the problem mentioned earlier of making the Rorschach evaluation more objective. Unlikely as it may seem, "general adjustment" is a rather simple thing to measure by the test. After all, we defined the term originally as having a "good" protocol and elaborated our psychological concepts after the fact, when we found that this definition worked out well in practice. To our own surprise, we found that the method we have developed for quick inspection of the protocol actually yielded a *numerical score* of impressive validity independent of our expert judgment.

What we did was to prepare a mimeographed check list of 30 items generally considered significant in Rorschach diagnosis. (Note these items remain in technical Rorschach terms--F per cent, color shock, CF FC, etc. They are not translated into judgments of behavior. The reader unfamiliar with the test must be content to assume that this abracadabra makes sense if it actually works.) Its original purpose was merely to provide a guide for *systematic* review of the whole personality as represented in the test, and a way of recording our findings quickly for future reference. We tried to omit nothing of general importance and also to include only the major points in each sector of evaluation—color, form, movement, shading, content, etc. Our method of recording was to enter a check against any item on the list where the protocol under consideration showed a marked deviation from the usual. Two or even three checks were entered when the deviation was very marked. Thus normal reaction to the appearance of color received no check, mild color shock one check, severe color shock 2 or 3 checks.

The rating discussed above is based upon a qualitative evaluation of all the data we were able to grasp in a short time, and it is more discriminating than the quantitative method now to be presented. Looking over our material, however, we found that simply *adding up the number of checks on the list for each student gave us a figure which corresponded well with the external criteria used to check the ratings.*

Sixty-four students out of 121 had 6 checks or less. None of these girls was markedly disturbed and only one had even a mild, temporary upset. Conversely, the group of 31 girls who had more than 10 checks included all but 2 of 19 students who showed fairly serious difficulties, academic or personal. In fact, only 4 of these girls did entirely satisfactory work and the descriptive comments of teachers suggested that none of them could be considered well adjusted.

It does not seem possible, by *counting*, to evaluate degree of disturbance among the 31 girls having more than 10 checks. To date, this finer discrimination can be made only by the judgment of the examiner based on more complicated analysis. Reduction of a group of 121 to 30 for more careful study is of great practical importance, however, especially since filling out the check list seems to require far less experience and skill than orthodox use of the Rorschach. A rough knowledge of the scoring system is sufficient.

In comparison with this method, we also tried out Dr Harrower-Erickson's criterion of "neurotic signs," equally objective and somewhat quicker. Students having not more than one "neurotic sign" kept out of trouble to the same degree as those having not more than 6 checks. No other discrimination could be made by the "signs," however. Girls with 2 "signs" had difficulties almost as often as those with 3 or 4. This finding is not surprising if the neurotic signs are actually a measure of overt psychoneurosis. Very few college students have the open symptomatology characteristic of diagnosed patients. Neurosis *narrowly defined* is by no means the only reason for failure to handle life situations effectively.

Reflection on the nature of the check list suggests that it works because, by design, it offers a *systematic* and *comprehensive* coverage of the resources of the personality. Adding up checks, therefore, becomes a meaningful procedure. Upward of 10 single checks scattered all over the lot actually mean a diffuse disturbance very likely to reduce the person's effectiveness—and very likely to be missed by the "neurotic signs." More serious difficulty in one or more sectors of adjustment, represented by double checks and a multiplication of checks in the same area, is reflected in a high score, *unless* all the other resources of the personality are functioning unusually well. A subject with marked difficulty in external relations will have a high number of checks in the color area. To keep the total number below 10 his handling of all other aspects of the test must be almost perfectly sound. An adequate score means that *other resources have been tested and found good.*

As a rule, there is good correspondence between the neurotic signs and the check list—as would be expected from the fact that the check list includes all the signs. Discrepancies are likely to be clinically significant—a measure of the fact that secondary factors in the personality are either contributing unduly to its inadequacy, or on the contrary are functioning so well that the subject can handle his difficulties effectively. Our material suggests that; for unselected groups, it is worth while to spend the small amount of extra time required for recording the supplementary data.

In passing, I would like to throw out the suggestion that such tests as the Bernreuter predict "adjustment" badly, not so much because they are questionnaires as because the questions they ask are neither comprehensive nor systematic. Some types of failure in adaptive mechanisms are overemphasized, others neglected. Statistical item analysis does not handle this problem at all unless the experimental "bad" group presents a single syndrome. These questionnaires have avowedly started with a list of symptoms and traits, not with an over-all concept of personality resources nor even clinical neurotic entities. Items have been retained when they occurred frequently in a heterogeneous "bad" group. If we are correct in ascribing the observed success of the check list to its systematically inclusive character, it seems plausible to account for the observed inadequacy of the Bernreuter by the unsystematic character of its construction. (The Bernreuter was given to the group of 225 girls here discussed. Its prediction of emotional difficulty was somewhat better than chance, but not much. Failures in prediction did not show the relationship to the ACE described above for the Rorschach—nor was the large group of adjusted girls clearly delimited.) A psychiatrist would

understand at once that a person may have few of the "frequent" neurotic symptoms and still be very neurotic, and that some quite adequate persons may be consistently on the introverted side—a trend which scores strong neurotic tendency on the Bernreuter.

Again, our experimental material does not permit sound generalization. It seems likely, however, that our adjustment rating succeeded beyond other attempts of the sort, not so much because it was based on ink blots, as because of the way personality data are handled in the Rorschach and especially in our check list. The survey of the personality is complex, systematic and comprehensive. A questionnaire constructed on similar principles might very well serve the same purpose.

In summary, then, our findings suggest that a measure of general adjustment can profitably be separated from capacity to deal with a particular situation such as academic work and recombined with measures of specific qualities for prediction of actual success in a given field. This method may well improve statistical prediction markedly in the majority of cases and isolate for intensive examination the small group where failures in prediction are frequent.

A numerical figure of good validity was obtained by adding up deviant items on a check list, thus providing a relatively objective means of using the Rorschach. The success of this check list is probably due to its systematic, comprehensive survey of personality resources. Such balanced comprehensiveness is proposed as a basic—and heretofore neglected—principle in the construction of adjustment inventories.

THE DETECTION OF PERSONALITY IMBALANCES

By Gardner Murphy
College of the City of New York, N. Y.

INTRODUCTION

The study of normal personality by experimental, clinical, and biographical methods has progressed on a broad front with such rapidity that it is perhaps more important to attempt a limitation and focusing than an exhaustive picture of the methods now available. The task of personnel selection, guidance, and training, which already confronts us on so vast a scale, will be surpassed by that colossal duty and opportunity that the returning soldiers will present. Millions upon millions of young adults and those nearing or at the prime of life will demand from us an intelligent evaluation of their backgrounds, abilities, and interests. They will expect from us more than rule of thumb, more than wisdom and learned talk. Instead of playing safe and asking merely that opportunity be provided to interview briefly and give suggestions to all such men, it is suggested that we boldly define the best which we might be able to achieve, with the sky as the limit.

I take it that there will be effective collaboration of medical men, psychiatrists, vocational and personnel workers, and clinical psychologists, and that we are concerned here, not with the enhancement of the vested interests of psychology, nor with any honorific concern for departmental lines, but simply with the question of tools that those might use whose chief training is psychological.

This will mean that psychosomatic problems will be considered here from a viewpoint quite different from that of the physician. The physician's training will frequently permit him to trace out in full clinical richness the psychosomatic elaborations of each problem, including the manner in which the organic system is reflected in one's attitude to oneself and the world, and the manner in which the attitude toward oneself and the world is reflected in the organic system. To be complete, a survey of such problems would require discussing physiological, biochemical, and other medical techniques. My problem, however, leaving such problems to the physician, is simply to explore techniques of investigating attitude toward self and world through experimental, biographical, and other available methods, so as to observe tendencies within the individual which lead into psychosomatic problems or which

appear whenever such problems are present in the organism. We are concerned with psychological methods for evaluating individual personality, leaving the further elaboration of the psychosomatic problem to Dr. Kubie and his collaborators.

One more distinction is perhaps needed: the psychologist must regard imbalance, defect, or disorder as a disturbance of interrelations, the study of such interrelations being an important phase of the normal, universal, and general psychology of personality. In order to be of any service in exceptional cases, we shall have to emphasize in our whole approach the systematic study of the normal person.

All serious research springs, of course, from clear questions which we put to nature, hypotheses so framed that there is an answer. Among the hypotheses which the individual psychologist would have in the back of his head as he works, are some that deal with the general needs and problems of the population that he confronts, as well as some that are concerned with what is most needed by a particular individual. And those who would plan or administer such a program as we now envisage would have to develop sharply defined hypotheses as to the kinds of results that can reasonably be expected from different kinds of procedures. They will need to have clearly defined hypotheses as to the time available for the individual client. There will need to be hypotheses as to the type of psychological personnel available, the type of training it needs as background, the manner of staff organization and of collaboration with medicine, vocational guidance, psychiatric social work, and other public services. There will need to be hypotheses regarding the relative importance of a research program and a guidance program, and, finally, there will have to be a clear mandate to the individual psychologist regarding the relation of his guidance function to the broad research function which the gathering of such data entails. None of these hypotheses can, of course, be discussed here, but it must be kept in mind that the value of any method or group of methods to be described will depend upon the exact program determined upon and the personnel trained and assigned the task.

I will now attempt to present a bird's-eye view of methods available for the study of imbalances in the adult personality in our culture, with a view to defining the personality problems in such fashion as to lead into fruitful psychosomatic investigation. I should make clear that the use of the whole array of methods mentioned would take far more time than will be available in the individual case. I assume that a

selection will ultimately be worked out comprising a few basic methods used with all subjects and, in addition, methods *specially chosen* to help in each individual case.

THE INTERVIEW

To establish friendly rapport, achieve a first impression of the individual, obtain a few salient background factors, and define the subject's attitude toward himself, his future, and our own possible help to him, we shall have to begin with an interview of an informal and leisurely type, encouraging the subject to talk, and, in no event, hurrying him or prematurely narrowing the conversation. The case history obtained in the subject's own most natural manner will be of the semi-standard type, with certain entries and checkings of items for all subjects and with qualitative notes added in each case. During this preliminary interview, the examiner will determine which of the long list of possible methods are likely to be most fruitful in this individual case, over and above those general methods which will be included for all subjects throughout the program.

PERCEPTUAL TESTS

While personality study twenty years ago was. to a considerable degree, dominated by analysis of behavior, as such, it is fair to say that our first problems today are likely to be problems as to how the individual perceives himself and his world. They are, in other words, perceptual problems. To ascertain the way in which his perceptions are loaded by his personality structure, we shall present, at the beginning, a series of projective tests with semi-structured materials. I suggest, first of all, the disc prepared by Skinner[1] and used by Shakow and Rosenzweig[2] under the name of tautophone, a disc consisting of meaningless sounds which, under experimental conditions, are interpreted by the subject in terms of meaningful word sequences. The subject is asked simply to indicate the words that he hears. This may be followed immediately in the visual sphere of experience by incomplete pictures of the type used by R. N. Sanford[3] or by indistinct pictures as used by Robert Levine,[4] the pictures being interpreted by the subject in terms of his interests and drives, i. e., autistically structured.

[1] **Skinner, B. F.** The verbal summator and a method for the study of latent speech. Jour. Psychol. 2: 71–107. 1936.

[2] **Shakow, D., & Rosenzweig, S.** The use of the tautophone ("verbal summator") as an auditory apperceptive test, for the study of personality. Character and Personality 8: 216–226. 1940.

[3] **Sanford, R. N.** The effect of abstinence from food upon imaginal processes: a preliminary experiment. Jour. Psychol. 2: 129–136. 1936.

[4] **Levine, R., Chein, I., & Murphy, G.** The relation of the intensity of a need to the amount of perceptual distortion: a preliminary report. Jour. Psychol. 13: 283–293. 1942.

The pictures are, however, not only perceived in accordance with the relative strengths of various drives, but may lead, if encouragement is given, to active fantasy which may be recorded. At a higher level of structure, we may use the Murray thematic apperception test,[5] evaluating responses to pictures in terms of the subject's identifications with the people portrayed, and in terms of contents and quality of phantasy. Throughout such "projective tests," the data should be recorded in terms of verbatim records which can later be analyzed and scored. In addition, the examiner should take special note of marked disturbances of the individual in performance, blocking in speech or action, exceptionally rich autistic elaboration, or constricted performance with inability to elaborate fantasies. If it is possible for the examiner to indicate in a few words his major hypotheses about the subject at this time for validation later on in the research, so much the better.

Of outstanding value is the Rorschach,[6] given as an individual test and scored in the formal fashion but with free use of a supplementary check list on such characteristics as verbal facility, volatility, rigidity, cooperation. Abundant opportunities for short cuts such as the Munroe inspection technique[7] are, of course, to be utilized when necessary.

EGO STRUCTURE

Whereas the methods just described aim primarily at the perceptual dispositions of the individual, we need now to focus more sharply on one special perceptual problem, namely, the way in which the individual looks upon himself. What sort of a person does he conceive himself to be, by what methods does he enhance his ego, by what methods does he defend it? Suggestive data already at hand in the first part of the examination can be supplemented here by free chain association, two or three verbal stimuli being given, each one starting the subject off for one minute. A concealed galvanometer may easily be used, the deflections being studied in conjunction with the apparent affective intensity of certain responses. This may be followed by studies of the subject's basic expressed interests, as, for example, in the Allport-Vernon study of values,[8] or by Eugene Lerner's test of ego blocking,[9] in which the individual exemplifies in motor perform-

[5] **Murray, H. A.** "Explorations in personality." 1938.
[6] **Rorschach, H.** "Psychodiagnostics" [English translation. 1942].
[7] **Munroe, R. L.** Inspection technique. Rorschach Exchange 5: 166–191. 1941
[8] **Allport, G. W., & Vernon, P. E.** A test for personal values. Jour. Abn. & Soc. Psychol. 26: 231–248. 1931.
[9] **Lerner, E., & Murphy, L. B.** (Ed.). Methods for the study of personality in young children. Monog. of the Soc. for Research in Child Devel. No. 3. 1941.

ance the way in which he confronts obstacles interposed by another person. Some general paper-and-pencil test of personality could be introduced at this point, perhaps the revised Chassell clinical inventory,[10] or the Willoughby revision of the Thurstone.[11, 12] There is no objection to the use of such schedules, and there is much good in them, provided that two limitations are faced: first, the total score will often be of little use to us in the present task, except as confirmation of data obtained under conditions where self-deception is less easy for the subject; second, care must be taken to note which items and groups of items fit meaningfully into other data.

An opportunity might be made now to subject the individual to direct strain or frustration, as by one of the Rosenzweig tests,[13] so constructed as to induce the sense of failure, taking note of his predilection for one or another method of handling his frustration, such as repression of his failures, blaming himself, and blaming the experimenter. With any of these, a concealed galvanometric measure may be taken. The Mittelmann-Wolff methods[14] of studying finger temperature under such stress would undoubtedly be of psychosomatic interest. Great importance attaches, I believe, to Else Frenkel-Brunswik's demonstration[15] of the ease with which subjects, in evaluating themselves, show a proneness to use one rather than another of the various available mechanisms. I have in mind especially rationalization, projection, displacement, repression, and reaction formation. Nothing would be more directly pertinent to psychosomatic problems than a notion of the individual's characteristic defense mechanisms in relation to the strains and costs they entail, as well as the strains and costs they may serve to prevent.

As tension develops, a pause and re-establishment of rapport is needed. The interview might be briefly resumed with an emphasis upon ascertaining the subject's *group memberships*, group loyalties, and areas of security. With whom does he identify, whom does he respect, for whom will he gladly make sacrifices? Check-list data and qualitative observations here will immediately be followed by brief schedules of social attitudes bringing out the individual's "we" feeling with other social groups. Does he identify in terms of his age, sex, race, religion, home town, in terms of family membership, or in what

---

[10] **Chassell, J.** Experience variables record: a clinical revision. Psychiatry 1: 1–8. 1938.
[11] **Willoughby, R. R.** Some properties of the Thurstone personality schedule and a suggested revision. Jour. Soc. Psychol. 3: 401–424. 1932.
[12] **Thurstone, L. L., & Thurstone, T. G.** A neurotic inventory. Jour. Soc. Psychol. 1: 3–30. 1930.
[13] **Murray, H. A.** "Explorations in personality." Pp. 472–491 and 585–599. 1938.
[14] **Mittelmann, B., & Wolff, H. G.** Affective states and skin temperature; experimental study of subjects with "cold hands" and Raynaud's syndrome. Psychosom. Med. 1: 271–292. 1939.
[15] **Frenkel-Brunswik, E.** Mechanisms of self-deception. Jour. Soc. Psychol. 10: 409–420. 1939.

terms? Tests of the Bogardus[16] type have proved quick and adequate methods for such purposes.

At this point, we should stop and take stock of what we think we have accomplished by gathering such information. Having undertaken first to get a view of his manner of looking at the world and himself, we have now come to the point of asking whether he can accept himself, what efforts he makes to see himself according to his own standardized schema, and *at what cost* he maintains this self-portrait. Does he maintain an adequate self-portrait at the cost of physiological wear and tear, through maintenance of high-tension level and an abundant use of such psychoanalytic mechanisms as stand out even for surface inspection, or does he oscillate between casual, relaxed self-acceptance and strenuous corrections of the picture when shame and guilt, inadequacy and inferiority force themselves upon him? A rich mine of data here upon the subject's conception of himself should be made available for the medical evaluator of the subject's psychosomatic problems.

MOTOR EXPRESSION

We have moved slowly over from the perceptual field through the centralizing ego functions to the problem of motor expression. As I see it, we are not concerned in this conference with motor skills or aptitudes as such, but we *are* concerned with the field of motor behavior and expressive movements because of the eloquence of such behavior in portraying broad personality problems. Grace, clumsiness, even speed, erratic tempo, slow warming up, the end spurt—these are examples of the fashion in which any motor performance betrays personality strengths or weaknesses. Here, I would suggest some form of pursuit meter with and without distraction, a simple finger maze test, the Howells test of persistence,[17] despite physical pain, in the execution of a task, and a rough gauge of the subject's tension level while undertaking a complex motor task, as, for example, by the Johnson-Duffy technique[18] of squeezing a rubber bulb with one hand while executing a task with the other. I should suggest adding, at this point, Eisenberg's procedure[19] for determining the sheer expansiveness or withdrawal tendency of the subject while carrying out motor tasks, with the hypothesis that freedom vs. constriction will show itself in the ac-

[16] See Bogardus, E. S. "Immigration and race attitudes." 1928.
[17] Howells, T. H. An experimental study of persistence. Jour. Abn. & Soc. Psychol. 28: 14–29. 1933.
[18] See Duffy, E. Tension and emotional factors in reaction. Genet. Psychol. Monog. 7, No. 1: 1930.
[19] Eisenberg, P. Expressive movements related to feeling of dominance. Arch. of Psychol. 211: 73. 1937.

tual utilization of space. The purpose throughout these motor tests is partly to get quantitative data on the blockage, clumsiness, variability of effort of the subject, and partly to permit further qualitative evaluations.

SUMMARY IMPRESSIONS

For a few moments after this battery of procedures, the interview may be continued, summarized, and concluded with a few words to the subject about his strong points, and about our hope that we can help him. During this period, there is an opportunity for the examiner to take special note of various additional general characteristics of his subject, a few of which can be entered upon a disguised check list and many more of which can be entered as the subject leaves the room. I have in mind such broad characteristics as ease vs. rigidity; compulsiveness vs. casualness; the range of intensity of individual anxieties, personal, social, sexual, economic, and the cost to the organism involved; general aggressiveness and areas of aggression; the tendency to throw the aggressiveness into fantasy form as against the tendency to carry it out directly, and, of course, the tendencies to reaction formation; the subject's general autism level, his tendency to easy wish fulfillment in perception or imagination, especially in relation to his picture of himself; his relation to his own body both in terms of his recounting of earlier illnesses or disabilities and in terms of his posture, gesture, and verbal self-reference. One thinks here especially of the problem of the returning soldier who is exuberant in his renewal of youth as he resumes civilian life, and of the man who is lost without the support of routine and authority upon which he has relied. I will leave to Dr. Kubie and others the discussion of this issue, but all of us need to keep it near the center of our *thinking*.

Now, putting aside the primary concern with the problem of imbalance, the subject's relation to the general personality types of normal people in our society should lead to final ratings on the subject's general and special abilities, his values and interests, his chief conditionings, his adaptability and educability, his psychophysiological adequacy, the areas in which he will most likely be adequate and secure, the life areas in which he would face the gravest hazard. All of these traits may be entered on check lists with a word or two of supplement here and there.

Again, to supplement and balance the emphasis upon pathology, the *resources, strengths* and *adequacies* of the individual should be fully defined. What does he have to live for and what are his tools for cop-

ing with the world? As Pierre Janet would put it, we have heard about his psychic liabilities. What are now his "psychic assets"? What is the "psychological income" which must be compared with his psychological expenditure? If the subject's strength is as clearly defined as is his weakness, this will lead us rapidly into a situational analysis of the subject, i. e., a consideration of his past and present environment, and especially of his future environment. We shall think not in terms of his absolute adequacy, but his adequacy relative to this or that world with which he must deal; a world in which he will be older, more experienced, supported by more and different kinds of people, as well as subjected to many specifiable and some unspecifiable types of strain. Instead of making a guess as to how such a person will come out five years from now, we shall, like the research people in the Cambridge-Somerville Youth Study, undertake to guess where he would come out in relation to various environmental problems, using such a gauge as a way of forcing ourselves to specify how he might be protected and aided in coping with each of these environments. Our problem, in other words, will be only half a problem of detecting his imbalances and trying to correct them directly. It will be half a question of gauging ways in which, through situational therapy, these imbalances may be prevented from making trouble or may even be turned into supports in his living.

COMMUNITY SERVICE

Having these data and interpretations and turning them over to medical men or administrators, or whoever bears final responsibility for the guidance of the individual, what will be our duty to the community, to which we are ultimately responsible and which we hope to serve in the broadest terms? For it is not only our obligation to focus clearly immediate needs of the individual; it is our obligation to build up an applied science which may protect and guide more and more effectively the thousands, yes millions, who so desperately need all the clear information that we can give to face the fiendish complexity of modern living. Just as every modern room is lighted, every subway train propelled by forces experimentally studied by Faraday a century ago, so we may hope that every human life will in time be more effectively lived because the present opportunity is used *not only for therapy but for basic scientific understanding.* While the individual doctor helps his patient, medical research eliminates yellow fever or diphtheria. Psychology must work with the same scientific concept

of its future. Now, for self-respecting research, it is important that a basic minimum of 15 or 20 methods be systematically applied to all comers in any large program, both to permit the study of interrelations and to give a base line for longitudinal studies of the same individuals after a considerable period. Beyond this base, however, individual methods may be used in so far as the psychologist thinks them valuable as a means of helping the individual. These data should also be worked up, *especially* in *longitudinal* form, but sampling difficulties will probably prevent their giving adequate data for group comparison purposes.

Finally, the point should be stressed that it has been an era of enormous proliferation of personality tests. It is now a question, not of dozens, but of hundreds of such tests. The chief problem is not to devise new methods but to develop these methods to a high level of adequacy. No personality test, except the Rorschach, has run the gauntlet and demonstrated its profound and far-reaching adequacy, but many of those in use may ultimately do so.

BROADER SOCIAL IMPLICATIONS

In conclusion, the purpose of focusing upon personality evaluation is not merely to help individuals and not merely to advance science; but to educate a democratic society in the importance of personality; to leaven the educational system as it learns to understand and to help individual children or adults; and to lay a psychological foundation in the community, so broad and so well understood, that, in the next generation, psychosomatic difficulties will be fewer, because sound applied psychology of personality will be taken for granted, just as sound physics, chemistry, and biology are taken for granted in relation to problems of health. One of the supreme tests of a democratic society is the question of its ability to make clear to its component individual members, through research and application, the dynamic principles upon which interpersonal problems can be solved; an important part of our job is to contribute to public understanding of the rationale of our methods and the meaning of our findings.

DISCUSSION OF THE PAPER

Prof. L. Joseph Stone (*Vassar College, Poughkeepsie, N. Y.*):

Professor Murphy has supplied us with the brilliant synthesis and program which he has led us to expect and demand of him. The program that he proposes is so thorough and suggests so full a use of the psychologist's armament and equipment that I shall not attempt any unnecessary comment or amplification. I prefer rather to approach the question of implementing his proposals.

I should like particularly to suggest that motion pictures could be of threefold service in such a plan. I shall omit from the present discussion any reference to the use of films for research purposes, not because research is not essential to an adequate plan, but because their use in research involves complications which it would take too long to follow out today. It seems to me that films could be of great significance in: (1) *training* the vastly expanded psychological personnel which such a program would necessitate; (2) as a *recording medium* in actual clinical use, providing objective records of clinical situations and at the same time economizing the time of the best trained clinicians in a way that I shall explain in a moment; and (3) as a vital means in meeting Professor Murphy's last suggestion to *educate the professional and lay public* to an understanding of the psychology of individuality.

I feel qualified to speak to the point on the basis of the fortunate opportunity I have had to participate in the program of film production of the Department of Child Study at Vassar College, a program under the direction of Dr. Mary S. Fisher, supported by grants from the Josiah Macy, Jr., Foundation and the General Education Board. Taking part in the production of a series of films (under the general title "Studies of Normal Personality Development") has given me an opportunity to become one of the members of the peculiar new species of "psychophotographers" or "photopsychologists." I can therefore speak in terms of filming techniques and possibilities as well as of the transmission of film-borne psychological concepts to students and lay audiences.

Filming is expensive on a small scale, but it is cheap when a large program is involved. The primary requisite for instituting so practical and significant a plan as Dr. Murphy has put before us to facilitate postwar adjustment is the speedy and efficient *training on a large scale* of competent psychological clinicians, clinical assistants, and technicians. The present wide use of training films in the Army and Navy, for instance, suggests that motion pictures provide an ideal means of combining the handling of large groups of students with the provision for each student of the necessary intimate acquaintance with materials and techniques that he is to use.

Moreover, we have found that at certain points in training, films are not merely substitutes for the direct experience that it may be difficult to afford a large group. Either in terms of actual learning about sample individual personalities or in connection with the training in the use of various projective techniques and other instruments, the film may be far superior to direct experience. By films I always mean films *with sound*. The inclusion of sound provides the opportunity to record a situation in its completeness and entirety. It also makes possible, by the addition of interpolated commentary, to point up what is to be observed at the moment that it is being observed—something that is particularly difficult in the observation or practice of actual clinical techniques. By the judicious use of such commentary and by careful selection and editing of the film, it is possible (for example, in demonstrating a specific technique such as an interveiw) (1) to: present a *wide range* of contrasting responses such as it would take a long time to observe in actual practice; (2) *select* the livest and clearest and most significant material rather than the haphazard run-of-the-mine material that would be seen in direct observation; (3) provide *common observation experience* for all the students in a large group to discuss; and (4) make possible the *repetition* of a clinical experience and the demonstration of fine points of behavior and attitude which may not have been observed at first.

May I remark incidentally that I feel that this point is the chief essential in the *objectification of clinical intuition*. The skilled clinician seeing and hearing again the behavior of a subject in a projective technique or in an interview could say "this and this and this made me decide thus and so about this person." He could point to the turn of speech, the sudden gesture, the covert glance that implied anxiety about certain problems in the patient's life history—or whatever significant trait or attribute was under discussion. What I am suggesting is that "clinical" intuition" refers to the clinician's accumulated experience that enables him to interpret swiftly and (sometimes) surely the "language of behavior"—Dr. Fisher's felicitous phrase—of his patients. The specific cues in behavior, intonation and gesture ordinarily may be too numerous, too slight, or too rapid for a

full account from memory. In training, however, where it is essential to teach students *to see,* the clinician could go back over the swift panorama of behavior in the situation as many times as necessary to indicate in detail the clues to the understanding of himself that the subject or patient provides in a well-chosen clinical situation. Carl Rogers' comments on the value of phonographic recordings of interviews for such purposes suggests how much more useful the complete record of the sound film would be. Much of the subjectivity and *sui generis* character of interpretation and conclusion that critics such as Macfarlane complain of, in the field of projective techniques, for example, could be overcome by the explicitness of this method.

Good, closely related discussion and study material must be planned around the film, but no other method can give large groups the benefit of direct, intimate, repeated and common observations as the basis for discussion.

The second suggestion that I wish to make—the use of automatic sound-film *recording* in some of the actual clinical situations—is tied up with the obvious point that, if any such large-scale programs as we are here considering are put into effect, the best trained personnel will have to be spread very thin. It will be necessary for many purposes to use the services of less thoroughly trained clinical assistants and technicians. I believe that preselected key portions of the administration by these assistants of various projective techniques and other instruments should be recorded with concealed microphone and camera. The chiefs of a service could then run through a screening of the responses of a number of subjects and make their judgments in a considerably reduced time. This might well reduce greatly the number of individuals, already examined by assistants, who would need to be called back for further time-consuming study by the chiefs of service.

Finally, let me say that the use of films for the *educational presentation* of a basic point of view toward personality—the third use I wish to suggest—is closest to the work which we are doing with our films, and I can speak of it with greater confidence. All that is required—and, of course, it is not very simple to achieve—is a clear conception of what is to be presented, a combination of imagination and accuracy in its presentation and adequate (not exorbitant) funds to make possible the translation into films. I should like to sound the warning that it is essential for the psychologist engaged in making such films to keep in close touch with every step of film production, and not simply to farm out the general plan to psychologically unsophisticated film experts

To achieve Dr Murphy's goal of educating a democratic society to the point where "a sound applied psychology of personality will be taken for granted," we need a realistic and effective program for psychological mass education. Few techniques can match the film for this purpose. To prepare a vast range of films from popular presentations with good insight (such as "Journey for Margaret") to more technical presentations for college use challenges the imagination of every scientist whose work touches the field of personality; challenges him to meet the obligation he bears to a democratic society; to translate his research and clinical insights into common knowledge.

Dr. Robert C. Challman (*Teachers College, Columbia University, New York, N. Y.*):

I am sure we would all agree that Dr. Murphy has done an excellent job in organizing the pertinent material around his topic. I have a few general suggestions and a few specific ones. First, if we plan to deal with young adults from widely differing cultural backgrounds, we must make sure that all our tests and other approaches are suitable for such a broad range. Second, the plan for the psychological study of these men must take into account their motivation for submitting to the various instruments and approaches. If we are dealing with volunteers and use a strong frustration situation as one of our approaches, we may find that many of our subjects may respond by a bodily and permanent "leaving the field." Third, as no mention was made of the detection of personality imbalances and inadequacies through intelligence, achievement, and aptitude tests, I assume Dr. Murphy deliberately omitted them in order to emphasize the tech-

niques which bore a more direct relation to personality. However, as all of us realize, much "personality data" can be gleaned from such tests.

As to specific suggestions, I would be inclined to omit a measure of galvanic reactivity. It seems to me that if the research of the last twenty years has demonstrated anything, it has demonstrated that we still do not know enough about the conditions underlying galvanic reflex to be able to assess its significance for personality. It would also appear that such techniques devised for children, such as the Lerner Ego-Blocking technique would not be applicable to adults unless extensive modifications were made in the tests.

It might also be fruitful to substitute the autobiography for Chassell's Experience Variables. Dr. Murray, of Harvard, found the autobiography of definite value in his work with college men, and the task itself is not too difficult for those with fourth-grade education or above. An adaptation of the Mooney Problem Check List might help to bring into focus the problems of the individual. A further development of "conflict stories" in which the individual is confronted with a story of a life problem that allows only two solutions, each one involving a denial of the other, might be used. Dr. Herbert Zucker found this approach very valuable with delinquent boys. In contrasting attachment to parents with that of attachment to an age mate, he told boys stories similar to the following "Johnny is called into the principal's office and told that something terrible has happened to his parents and that they want him to come home immediately. While he is in the office the phone rings and he is informed that his best friend has been in a serious accident and wants him to come to the hospital right away. Where does Johnny go?"

The use of self-ratings on abilities of various kinds might be used as a means of obtaining the individual's own evaluation of himself rather than as valid indicators of actual abilities.

Finally, I would like to mention the level of aspiration technique Through this method, a worthwhile indication of self-confidence can be obtained, and a direct estimation of both the intensity of the individual's reaction to failure and the way in which he handles failure can be made

Dr. Eugene Lerner (*Sarah Lawrence College, Bronxville, N. Y.*):

The topic so suggestively treated by Dr. Murphy could relate to three problems: (1) selection of military personnel for immediate war purposes, (2) readjustment of both healthy and more or less incapacitated military personnel to postwar civilian life in the United States, and (3) rehabilitation in various European countries as part of world-wide psychological reconstruction in general.

Distinction ought to be made between what may be called preventive personnel selection and palliative personnel selection Though we are today of necessity focusing on the latter, simply because we as yet lack the requisite longitudinal data on a nation-wide scale, it seems important to emphasize what would be needed for truly preventive personnel selection. The two methods are clearly interdependent. Mr. Frank suggests that two points are important: (1) how did the subject or candidate get "that way," and (2) can he take it, as he is constituted now? It seems to me that for really adequate personnel selection the two approaches must be dynamically connected, in order that we may answer the basic functional question· how did he get this way so that he can now take it in the manner in which he can; how did he get where he is right now in his capacity for taking it?

Preventive personnel selection would presuppose continuous longitudinal inspections on a nationd-wide scale in terms of personality development in the first fourteen or eighteen years of the life span. Even if we cannot hope for the kind of comprehensive data secured in the California studies under Jean McFarlane, it may not be unrealistic to urge and plan for the systematic collection of minimal personality data on all or most children from nursery school or kindergarten on through high school. This would call for large-scale, nation-wide pretesting, testing and retesting—at age 5 or 6 years, at 9, 12 and 15 or 16 years (calling for 4 tests per subject, at 3-year intervals as a minimum). In addition to individual Binet-Termans at 6 and 12 years, group tests of intelligence, aptitudes and per-

sonality may be secured at 9 and 15 years, if not all 4 times. With the development of group-test methods of short inspection in the use of the Rorschach, this dynamic personality test may also be used at least twice—at 9 and 15 years, if not more often. Controlled 10- or 15-minute guidance interviews may be secured similarly, at least twice, at 6-year intervals. From age 9 years on, paper-and-pencil forms of simplified personality inventories may also be considered at 3- and 6-year intervals, including suitably adapted tests of moral judgment and morale—to gauge gross ratios of social-emotional maturation in terms of social personality trends. All such data would constitute a child's school record, pretty much as the properly "academic" part of it does at present. It would call for immediate and rapid extension of the use of school psychologists, visiting teachers and school psychiatrists, including bureaus of child guidance and vocational guidance in school systems throughout the country. It would call for greatly extended facilities for training the required psychological personnel, some centralization of scoring and interpreting facilities on a mass-production basis and so on. Considering what we are willing to spend on military equipment in time of war, the funds needed would seem relatively negligible—especially since the information thus secured would be vitally relevant to any future war effort, in terms of both military personnel selection and postwar rehabilitation programs. At least, it seems important to emphasize that in the absence of such comprehensive, if still minimal, developmental or longitudinal data, all palliative efforts are doomed to being severely limited and rather liable to serious errors. Whether for personnel selection or any other important purpose, personality cannot possibly be gauged reliably and with lifelike meaningfulness without a rock-bottom minimum of the total developmental picture or gross rate of social-emotional maturation A purely palliative picture, based entirely on test data *as at present*, may call for diametrically opposite diagnostic and prognostic evaluations—precisely in the light of such developmental test and observational data.

If nothing else, we now ought to be clear about what to expect and what not to expect from the palliative programs, however carefully planned. On this basis, I would recommend as additions to Dr. Murphy's suggestive batteries the use of standard intelligence tests, especially in terms of relevant "incidental" reactions to the test. Such seemingly tangential responses can be indicative of important attitudes, personality needs or characteristics. Also some simple job-tests, especially as used in the study of industrial accident-proneness by industrial psychologists, may be helpful in rounding out not only the aptitude profile but the total personality picture as well. Insofar as possible, standardized, if brief, autobiographic sketches may be obtained on group-test basis, in addition to group Rorschachs on short inspection basis. Thematic analysis may be centralized on a mass-production basis, under the leadership of the Harvard Psychological Clinic and similar clinical-experimental centers of personality study. At least the more serious extremes may be recognized and selected for more intensive follow-up examinations in such manner. A simple form of test for gauging rumor-proneness may also be used on a paper-and-pencil group test basis, in terms of Gordon Allport's and Robert Knapp's work. Such projectively gauged data may be more helpful and may take the place of lifelike behavior sampling in terms of miniature life situations—in comparison, say, with the Bernreuter self-sufficiency scores and similar techniques suggested by Dr. Murphy. Proneness in terms of anxiety rumors, hate rumors or day-dream rumors may reveal lifelike clues to such problems as management of aggression and anxiety needs, objectivity, rigidity, emotional stability, optimism-pessimism, etc. Paper-and-pencil "audience reactions" of standardized briefness, in response to standard film shorts may similarly be considered here. Retesting on the whole battery or parts of it, at intervals of six months or a year, would be necessary for more than static personnel selection, even on a palliative basis. The changes or "consistencies" thus revealed would be necessary for gauging continuing personality changes and checking minimal prognoses made at the time of first examination. The effectiveness of rehabilitation and treatment programs could not otherwise be verified and improved.

These approaches would be useful not only for postwar rehabilitation and planned demobilization of American soldiers, including those wounded or trauma-

tized, but also for postwar reconstruction of Nazi-indoctrinated and Nazi-traumatized countries abroad, including soldiers and civilians, children and adults, on a however limited, partial "sampling" basis.

Dr. A. H. Maslow (*Brooklyn College, Brooklyn, New York, N. Y.*):

Dr. Murphy has presented us with a vast array of possible tests that can be used in the detection of various trends of character and maladjustment, and therefore in the screening of draftees. It now remains to discuss the practical question of just which tests to use and when to use them. Certainly we cannot use all these hundreds of tests. I will say briefly that my reaction based on my own experience has been that if I had to use any single test of the many that are available, I would rely most upon the Rorschach test. If I could add to it, I would add simply any one of the standardized intelligence tests. My impression is that these two tests are probably more useful than any other brief combination.

However, this is not a complete answer either, because, when I say Rorschach test, I mean Rorschach test as interpreted by a competent person and of these there are very few. The Rorschach test has the great disadvantage, as compared with other personality tests, that it is not foolproof. It can be used badly by a poor psychologist and can do a great deal of damage. This means that we must use it only when we are certain that our interpretations are sound and expert

I must report to you the unpleasant fact that I know of several persons who are *supposed* to be Rorschach testers, who come out with the most remarkably and completely incorrect diagnoses. Such psychologists are dangerous because they are cloaked by the known reliability and validity of the Rorschach test and they forget, as do those who trust them, that this reliability and validity of the Rorschach test is not of the test itself but of the test in the hands of well-trained and competent investigators. In other words, the Rorschach test is definitely not a foolproof test. I know of no way in which it can be made so. There will always be untrained people who will nevertheless consider themselves to be adequately trained and we have no social techniques with which to prevent them from considering themselves in this light.

I do not mean to imply by this foregoing recommendation that I have any lack of faith in the so-called paper-and-pencil test. I am a little impatient with the derogation of these tests that is so common. It is true that these tests have been used unwisely, perhaps in more cases than those in which they have been used well, but this is true for practically any kind of scientific instrument. I have just pointed out that the Rorschach test, which is certainly valid, may also be used unwisely. I do not consider this something against the test Even the paper-and-pencil tests that are very frequently condemned as atomistic by various psychologists need not be, if they are used by a person who knows how to use them. I would use the analogy here with the surgical scalpel. In the hands of a skilled surgeon it is a wonderful instrument; in the hands of the layman it is dangerous. I think that the parallel applies very well to personality tests. If a person approaches the study of personality with a holistic attitude, then he can use wisely and practically any datum, however obtained, simply by putting it into its proper place in the total context. It is not his fault that other people may use precisely the same data as simply a kind of arithmetical sum of the separate traits or as isolated bits of information.

Finally, I wish to say a word about a basic question that is, I am afraid, embarrassing for most of us no matter what our training. This is the question of the interview. I myself must confess to relying more upon my interview than upon any tests or combination of tests, and still at the same time I must recognize various objections to this procedure. I suppose we would all agree that an experienced and well-trained psychiatrist or psychoanalyst, or psychologist, can detect more in a short half-hour or hour interview than can be detected by a whole squadron of mechanically minded testers. But at the same time, this person cannot ordinarily tell exactly how he derived this information nor can he readily teach others how to do it. It is my experience when I have asked these good interviewers just why they came to a certain conclusion, that if they were completely honest, they would say "I do not know," and that others would give

me various statements that I have generally considered to be pure rationalizations. Because of this I have come to the conclusion that the good interviewers get good results in spite of not knowing how they get them. I refer you to the work of Dr. Werner Wolff on the subject of intuition which will certainly throw a good deal of light on this phenomenon which I have just mentioned.

As things stand now, we do not have a situation in which we can simply train people to do a certain job. Certain people, I believe, can never be trained to be good interviewers simply because they must be certain *kinds* of people themselves. So the situation arises in which most of us are willing to make recommendations for other people that we do not apply to ourselves, simply, I suppose we must confess, because we trust ourselves more than we do other people en masse. Perhaps for the sake of avoiding the misdeeds of those people who are sure they are good interviewers and who actually are not, it might be well for all of us to agree, no matter what our faith in ourselves, to supplement our interviews by various other more mechanical, more objective techniques if only to check our interview impressions.

Finally, I cannot resist saying before this audience that I think that ultimately all the personality tests that we now have available will become invalid for the simple reason that practically all of them can be faked. Malingering is *possible* in all of them even though it may also be difficult. My own feeling is that ultimately we can and *should* develop non-malingerable personality tests and these I feel will be mostly physiological, or if you prefer, psychosomatic. The various indices of tension, of autonomic reaction, and the like, cannot be faked except with the utmost difficulty. It is here that I myself would wish to look for the ultimate answer to the diagnosis of personality. When I first had this thought years ago, and thought of working in this direction, it was so wild an idea that I didn't dare speak about it to anybody. In the last year or two, however, it has become quite acceptable and, as a matter of fact, one could make a clear case on this very day for a battery of physiological tests which could do the job with only a little more research. I think as it stands today, such a battery would be exceedingly unwieldy and take far more time and also have rather questionable results, but it is already clear that we are going in this direction and, furthermore, that we can go farther in this direction.

I myself am now working on the standardization of a paper-and-pencil test that is based directly upon the fact that there are somatic expressions of personality maladjustments. This test is in actuality a concealed test of emotional security and it is presented in such a way that no person yet has suspected its ulterior motive. Perhaps this is another possible answer to the question of faking tests. The usual personality test tells very clearly what it is seeking to find out. For the person who does not want this to be known, it is very obvious that he can slant his answers in a direction that is to his own benefit.

Dr. Morris Krugman (*Bureau of Child Guidance, New York, N. Y.*):

As one struggling daily with the problem of early detection of personality disorders, I agree wholeheartedly with the major aspects of personality and their methods of evaluation, as outlined by Dr. Murphy: the subject's way of seeing himself and the world, the ego structure, behavior and expressive movements, and psychoanalytic mechanisms. The only exception, perhaps, to wholehearted agreement, is the last grouping, which pervades and overlaps the others. Since the psychoanalytic mechanisms are, or have been, the province of the psychiatrist, some would argue that the psychologist should limit himself to intellectual aspects of personality, leaving the affective to the psychiatrist. Perhaps this is a question a psychologist should not bring up. Actually, this is not an embarrassing subject for a psychologist. His methods of attack on the problems of personality evaluation are so different from that of the psychiatrist that many psychiatrists seek the psychologist's evaluation both for the leads it affords him, and for the confirmation he so often desires. Furthermore, as clinical service and personnel proceduces are organized today, the psychiatrist is apt to deal only with the more serious personality aberrations, while the psychologist frequently deals with the general run of humanity, often acting as the screening agent and passing on to

the psychiatrist those individuals who are, or seem to be, seriously disturbed. At least, that is the way it functions in the organization with which I happen to be connected, where psychiatrists and psychologists work side by side.

Referring specifically to the techniques mentioned by Dr. Murphy, their present status can be roughly classified into three groupings: (1) those that are useful today, having demonstrated their clinical validity; (2) those that seem to have possibilities for usefulness, but have not as yet demonstrated their clinical validity, so that they cannot, as yet, be used as clinical tools; and (3) those that have proven of little or no value, or which do not add much to other techniques in use. Very few are in the first group. As Dr. Murphy has indicated, the Rorschach is outstanding among these. The autobiography, used as Murray did in his studies of personality, is another. These are useful because they are global in nature, because they permit the subject wide scope to project his personality, and because there are no "right" or "wrong" answers which an intelligent subject can control. Another important consideration with these two techniques is that, in the words of Dr. Murphy, "final judgment must depend upon clinical insight and integration of data." One of the weaknesses of some of the other techniques listed is that they rely solely upon "objective" scores, and not enough on clinical insight. It is my guess that personality will not be measured by one or several numerical scores without the use of the clinical insight of the examiner.

The second group, that which has demonstrated possibilities, but which will require much research before the techniques can become clinically useful, includes, in the main, the dynamic approaches, the global attacks on the problem, and the projective techniques. The Thematic Apperception Test, the Cloud Pictures, and the various association tests are examples. The possibilities in this group are limitless, and Dr. Murphy's suggestion that present known tests be developed rather than new ones devised, is sound.

The third group contains the bulk of the pencil-and-paper personality tests, and the tests which purport to measure single traits. Most of these devices do not stand up in tests of clinical validity, and the reasons are obvious. The pencil-and-paper tests of personality, most of which contain direct questions about attitudes and feelings, need not be honestly answered by the subject. The single trait tests fall down because traits are not static—they vary so much under different conditions that, in order to measure a trait adequately, it must be measured in most of these circumstances, and this is almost never done in practice. The introversion–extraversion scales, the ascendance–submission tests, the "neurotic" inventories, and the so-called "personality" or "emotions" scales fall in the two categories discussed under the third group.

Under Dr. Murphy's discussion of psychosomatics there is one possible omission: the electroencephalogram, which seems to possess many possibilities for personality evaluation. In addition to diagnosing organic and certain psychotic conditions, there is the likelihood that some types of personality disturbances may show characteristic wave patterns. Electroencephalography is another area in which studies are now being engaged in by the psychiatrists, psychologists, neurologists, and physiologists, working in a manner suggested by Dr. Frank in his remarks.

In addition to the techniques thus far discussed, we sometimes forget about the possibilities contained in the straight psychometric battery. In a child guidance clinic, for example, a psychologist sees a child for three or four hours in individual examinations. Added to observation and rating possibilities, a subject often yields clues about attitudes, feelings, relationships, etc., which help build a personality picture. Furthermore, test patterns, with which some research has been done, although not enough, give valuable clues. Test discrepancies, peculiar or atypical responses, differences of approach, changes in mood and tempo with different tests, expressive movements, comments, reactions to content, types of associations, display of energy or lack of energy, etc., all yield important information about personality. Performance tests are particularly fruitful, since the test atmosphere is much freer than during verbal testing.

These are only a few of the possibilities for the early detection of personality disturbances. They may not yield a diagnosis, but they do yield important data from which diagnosis can be made, or, at least, facilitated.

THE DETECTION OF POTENTIAL PSYCHOSOMATIC BREAKDOWNS IN THE SELECTION OF MEN FOR THE ARMED SERVICES

By Lawrence S. Kubie

College of Physicians and Surgeons, Columbia University, New York, N. Y.

INTRODUCTION

The weeding out of potential psychosomatic casualties before they break is a problem in practical organization as well as in medical science. Therefore, this paper will have to deal with both topics.

During the training period, psychosomatic disturbances constitute a heavier load in our military hospitals than do all other disabling conditions combined. To this initial load must be added a later crop arising in men who have been wounded or at least exposed to combat conditions, and in whom the slow evolution of a traumatic war neurosis finally focuses on a persistent psychosomatic complaint. This later group emphasizes the importance of immediate treatment of the traumatic war neurosis during the acute phase. It is of further importance from the point of view of the Veterans Bureau and of pension or compensation legislation, because unwise policies in this respect greatly increase the tendency toward the development of these delayed psychosomatic fixations From the point of view of the selective process, however, it is the *training camp casualty* that is of primary importance; and we shall confine our discussion, therefore, to the problem of the early recognition of the registrant who is likely to develop a psychosomatic disturbance during his preliminary military training.

The volume of this problem is staggering. It is encountered in every ward in every military hospital. From all come the same story: namely, that between 50 and 60 per cent of all patients are neuropsychiatric disabilities with a psychosomatic component. An orthopedic surgeon who is a consultant in the Air Force said recently that in two years of service, during which he has seen every supposed orthopedic case in a large area, not one turned out to be orthopedic; all were psychosomatic.

This situation obtains in every branch of the service, on general or special medical wards, on surgical wards, and in the venereal disease stations. It is important to keep this fact in mind when reading army

or navy statistics, because with a misplaced sense of chivalry, on discharge from the services, these patients are given a camouflage organic diagnosis. This not only renders our official medical statistics dishonest and valueless; it also does direct injury to the individual patient, because it fixes his attention on a nonexistent organic illness, and puts into the hands of his neurosis an argument to block all subsequent efforts at therapy. Thereafter, he can always point to the "organic" diagnosis in his discharge papers.

Thus, we find ourselves confronted by an astounding situation. By conservative estimate, of the admissions to training camp and base hospitals (omitting actual battle casualties), about 60 per cent of army and navy medicine is psychosomatic. Yet not more than 1 to 2 per cent of the medical personnel has even a rudimentary psychiatric training. Obviously, the patient cannot be treated fully in the service. How then can the potential psychosomatic breakdowns be recognized and kept out of the services? Before attempting to answer this question, we must consider what the psychiatrist looks for when he is trying to spot these conditions when they are fully developed. This will give us a hint as how to anticipate them. The inquiry must answer two questions:

1. What are the processes by which disturbances generated at the psychological level of the body's experience are translated into somatic dysfunction?

2. What clinical combinations do we find of the various physiological disturbances and the different psychopathological settings?

The processes by which psychologically generated tension states can be discharged wholly or in part through some disturbance of bodily function range themselves in a series of increasing physiological complexity.

1. There are certain well-known and banal physiological phenomena which are customarily looked upon as inevitable concomitants of conscious emotional states, such as sweating, blushing, palpitation, dyspnoea, and the like. Yet the association of these symptomatic bodily states with conscious emotional processes should not be taken for granted merely because they frequently occur together. A more careful consideration makes it clear that these bodily phenomena are not intrinsic components of any emotional state as such On the contrary, their primary physiological significance and purpose is to adjust the body to changing states of activity. They occur as necessary components of the total physiological pattern of fighting, attacking and escaping, etc.—all patterns of behavior which arise as a consequence of

emotion. Thus, when such bodily states as these occur "uselessly," that is, when they occur as concomitants of an emotional state, but without any change in total bodily activity, they lose their homoeostatic function. Indeed, unless changed bodily states demand them, they actually disturb homoeostasis. They become more than useless; they become destructive. This constitutes the first step in the dissociation of a physiological function from its primary, elemental physiological significance. In other words, the association between emotional states and such bodily phenomena as sweating, shivering, shaking, blushing, palpitation, dyspnoea, and the like, is the first of an ordered sequence of bodily substitutions, which leads ultimately to the most complex of the psychosomatic disturbances. If this is true, then, in any effort to predict psychosomatic casualties, the first task that confronts us is to ascertain how easily and under what psychological conditions any individual develops tachycardia, dyspnoea, sweating, blushing, shivering, and the like.

The method by which this can be done on large masses of men without great loss of time is something to which we will return later.

2. The next step in this process of dissociation is when these same commonplace physiological phenomena occur either apart from any concurrent conscious emotional states, or in a setting of weakly felt emotions. Here they are frequently spoken of as "emotional equivalents." It is a current fallacy to speak of such pallid or absent emotions as "too weak" to account for the accompanying physiological phenomenon. This linguistic short-cut is unfortunate, because it tends to perpetuate the illusion that these physiological changes have their origin in the emotional state as such; whereas, as we have already pointed out, their direct quantitative relationship is with changes in bodily activity, and the link to the emotional state is only through a secondary conditioning. Therefore, the term "Emotional Equivalent" is another misleading oversimplification.

It is important to emphasize the fact that both of these disturbances in the subtle equilibrium of the body are true psychosomatic phenomena, in the sense that through them tensions generated in the psychological sphere are discharged partially or wholly. In both of these cases, the physiological mechanism employed is one that is frequently associated with emotional states. For this reason, they have erroneously come to be accepted as intrinsic and inevitable parts of the emotional process and thus their significance in the hierarchy of psychosomatic phenomena is usually overlooked.

The methods of eliciting data concerning this second type of psychosomatic disturbance will be discussed below.

3. Next in complexity is the phenomenon in which psychological tension is discharged through a physiological mechanism that has no apparent emotional connotation at all. Examples of this are the "hysterical" anaesthesiae, paralyses, speech disturbances, trances, disturbances in orientation, and the like. Here the disturbed function has no direct relationship either to emotions or to bodily effort. Instead, the orientation of the individual toward the outside world and his ability to communicate with others are disturbed; i. e., the organs of the ego, the instruments by which the ego experiences the outside world and communicates with it. Characteristically, these dysfunctions occur in a setting in which conscious emotions are relatively in abeyance.

Clinically, this group is usually spoken of as the "conversion hysterias," and the process by which psychological tensions are discharged through the organs of the ego is spoken of as "conversion," reserving the term for this group, and not using it either for the simpler or for the more complex types of psychosomatic interrelationship. There may be some doubt as to the wisdom of setting this particular group apart by using a special name for the process, since this would carry the implication that the process of somatization at this level bore no relationship to the processes of somatization of simpler or of more complex nature.

4. Still more complex processes of somatization are seen when the physiological disturbances involve organs in the interior of the body, hidden away from the direct knowledge or eyes of the patient. Concerning the functions of some of these organs (such as the stomach, intestinal tract, or bladder), the patient has some knowledge, since from them he can receive localized sensory impressions. Of others, however, such as the bone marrow, spleen, or meningeal vessels, he has no knowledge at all. These disorders comprise what are usually spoken of as the "Organ Neuroses."

Thus it is seen that the processes by which tension, generated on the psychological level of the body's experience, can be translated into physiological disturbances, can be arranged in a continuous series from the simple to the most complex, varying at the same time from those that are most closely linked to conscious emotional states to those that ordinarily seem to be purely apperceptive, somatomuscular, or intellectual, and also from the most completely conscious to those most remote from consciousness. As we have already observed, it has been traditional in medicine and psychiatry to give different names to the

various stages of this continuous series. This has the didactic value of differentiating the different degrees of complexity sharply from one another. On the other hand, unless there is some over-all name for all, one is likely to forget that they are part of a continuous series of phenomena and to look upon them as fundamentally different processes. Probably, the best name to use is one that is most descriptive, least connected with any specific phenomena by past associations, and therefore least controversial. For such an all-inclusive name, we propose the term "somatization," using it as the name of any process by which a partial discharge of psychological tensions can occur through the somatic representation either of the emotion itself, or of the external relationships involved, or of the conflict out of which the tension arose, or, finally, of the instinctual functions which gave rise to the conflict.

We shall see that the recognition of these four basic types of somatization is only the first half of our nosological problems. Nevertheless, it is convenient, at this point, to stop and consider how, under wartime conditions, it might be possible to assemble information of this nature about men who are to be evaluated for duty with the armed forces

METHODS OF ASSEMBLING INFORMATION FOR PURPOSES OF EVALUATION OF MANPOWER

One would naturally think first of the possibility of using standardized physiological tests. Theoretically, tests might be evolved which would indicate quantitatively in any individual the susceptibility of each organ system to conscious and unconscious emotional influences. Unfortunately, however, the fact of the matter is that such tests do not exist. Nor is it entirely certain that such tests can ever be devised.

In the first place, since no human beings are equally sensitive to all affective stimuli or, in all phases of human life, the test methods for setting off emotional reactions would either have to be so varied as to touch off all possible sensitive points, or else, before the test, each subject would have to be studied sufficiently to discover in what areas he was sensitive, so that the tests could be directed toward those particular grooves. Clearly, this is not a practical plan when dealing in terms of millions of men.

Secondly, while under laboratory conditions it is relatively easy to induce acute emotional stresses, it is difficult if not impossible to set up experiments which would simulate or reproduce the long-continued chronic emotional strains to which men in the armed forces are sub-

jected. Indeed, in many respects, the laboratory with its implications of security and protection can hardly bring to bear on the individuals to be tested any emotional forces that are true equivalents of those that confront men in battle. For these reasons, quantitative studies of physiological responses to experimentally induced emotional states may not prove to be as promising a field of experimental investigation as a casual consideration might lead one to hope.

If, on the other hand, one attacks the problem of physiological tests from a different angle, and subjects the individual to physiological stimuli, one may readily secure constant and reproducible data. In turn, this may make it possible to classify individuals according to the degree of their physiological responses. On *a priori* grounds, it is reasonable to expect that an individual who is hyperresponsive to a physiological stimulus might also be hyperreactive in responses to psychic stimuli. However, this is not necessarily true for all, so that in each case it would have to be tested. On the other hand, negative data would seem to be wholly lacking in value. That is to say, it need not follow that individuals who are not hyperreactive to physiological stimuli would necessarily be similarly stable in their responses to psychic stimuli. Therefore, such tests might be expected at most to weed out only a small group of positives (individuals who were hyperresponsive to both), while allowing another group of potential psychosomatic casualties to go through unrecognized; namely, those who are not hyperresponsive to physiological stimuli, but who are hyperreactive when under the influence of psychic stresses.

Obviously, therefore, any physiological tests will have to be standardized against psychological factors. And the only wholly valid way of doing this would be by checking physiological tests against both the past and the future. That is to say, the subsequent medical histories of tested individuals should be followed throughout their period of service in the armed forces, both under combat conditions and afterward. At the same time, a careful past physiological history of each such individual should be recorded, so that the physiological tests could be correlated with all past information as well. In this way, it would be possible to estimate the usefulness of such tests, when applied at the moment an individual stands between the past stresses and strains of civilian life and the future uncertainties that await him in the army or navy.

It will be clear at once that this is a research project. It is not one that can be tried on ten million men. But it could well be tried on

unselected samplings of those ten millions of men, picking the sample groups according to age, types of service, and the like.

Unfortunately, in this situation, a short-cut that can be used in the validation of certain other tests is not applicable. As Dr. Fremont-Smith has pointed out, there is a method of speeding up the process of validating many selective tests. Instead of watching the later course of the tested individuals over months or many years, an immediate partial validation can be made by subjecting to the tests individuals who have passed unscathed through severe combat experiences and by testing a contrast group of individuals who have broken under the same stresses Unfortunately, in this particular type of disorder, this short-cut is not usable, because once a man's physiological mechanisms have decompensated, any tests of those functions will measure only the extent of the decompensation, and will give no indication as to what the same tests might have shown before his breakdown.

Thus, we are drawn inescapably to the conclusion that, whether for the validation of physiological tests, or for a direct evaluation of individual registrants, a careful psychosomatic history cannot be dispensed with. It is a truism that a diagnostician is only as good as the history that he takes. In the field of psychosomatic medicine more than in any other, an accurate evaluation is impossible without an anamnesis.

Yet no system of induction can provide time for a slow and painstaking questioning of each potential soldier separately, either by a physician or even by trained clerical aids. Ten million men in the armed forces mean twenty million men to be examined and twenty million histories to be gathered Clearly, if this is to be done, some short-cut must be devised.

For every medical history, its form and the method by which it is to be gathered, varies with the purposes it is to serve The history that is gathered for therapeutic purposes cannot be used unchanged for the selection of troops. An entirely different methodology is needed. So obvious is this, that it is a disgrace that neither the army nor the navy has ever realized that special methods of medical history-taking must be developed for use in connection with the selection of troops, nor that special personnel must be trained in their use.

I wish to digress a moment to say that there is a reason for this shocking and costly oversight. The reason is that the taking of medical histories is a natural function of physicians, and that physicians of the armed services, like the physicians of civilian life, are concerned primarily with therapy. Even those who, in times of peace, are at-

tached to recruiting offices serve there only temporarily and look upon this as an unimportant and unpleasant temporary chore. This state of affairs will continue until a wholly separate division is built up, both in the army and in the navy, consisting of physicians, psychologists, psychiatric social workers, penologists, and statisticians, whose exclusive function through peace and war will be to work out better methods of medical selection, to train a specialized personnel in their use, and to supervise their administration. We shall return to this organizational problem below.

To be useful, history methods must meet certain obvious practical needs:

1. They must be formulated in simple questions that can be answered Yes or No by the registrant himself with little or no help.

2. They must be arranged on a form that can be read and scored rapidly, perhaps by mechanical devices.

3. They must be asked in such a way as to give no lead as to which answer indicates sickness or health.

4. They must be drafted in such a way that the eye of the examining military physician can pick up significant answers instantly as he runs down the side of the page.

Such forms and methods are quite possible to work out. Illustrative approximations will be presented in a moment It should be kept in mind that, in the process of inducting soldiers and sailors into the army and navy during an emergency, the registrants stand around for hours, awaiting their turn to be questioned and examined. At present, these hours are wasted, when they could be used for filling out just such forms as these. Trained civilian aids, circulating among the registrants, could assist those whose knowledge of reading and writing was so limited as to hamper in any way their use of the forms. Such a simple device as this has proved most useful in Canada.

It is also possible to have similar or identical forms filled out by family physicians, or by reputable outpatient clinics or hospitals. The forms could then be brought or sent to the induction station, for scoring.

Before drawing any conclusions, however, it will be necessary to evaluate every element in the history form by checking each of them against the future physical and mental health of men in the services.

The material gathered from such questionnaires falls into four clinical groups. These groups represent four different ways in which the body could be involved in the process of somatization. At the same time, the four groups correspond to four basic physiological and psychological constellations.

External Somatizations, involving the external structure of the body which relate the individual directly to his environment, to wit: striated and skeletal musculature, secretory and vascular structures of the skin, organs of speech, distance and skin receptors, organs of kinaesthetic sensation and of equilibrium. These structures together subserve the external relationships of the individual, his orientation in space and perhaps in time, his ability to communicate with others, the sensory impressions that reach him from the external world, and his conscious orientative faculties. The process of external somatization is what has been called "hysterical conversion."

Internal Somatizations, involving internal organs such as heart, blood-vessels, respiratory organs, gastrointestinal tract, glandular

Questionnaire 1

If you get upset in any way (that is, happy, gloomy, angry, afraid, tense, amused), do any of these things happen to you?	No	Yes
Does your heart beat fast?		
Does your heart pound?		
Does your heart pound in your head?		
Does your heart pound in your chest?		
Do you breathe fast?		
Do you sweat all over?		
Do you sweat in your face?		
Do you sweat in your hands?		
Do you feel hot?		
Do you feel cold?		
Do you get gooseflesh?		
Do you shiver?		
Does your skin break out?		
Do your eyes water?		
Do you cry?		
Does your nose run?		
Does your mouth get dry?		
Do you have to urinate?		
Do you have to have a bowel movement?		
Do you feel sick at your stomach?		
Do you get hungry?		
Do you get constipated?		
Do you get a stomach ache?		
Do you have cramps?		
Do you have pain any place? (Specify where)		
Do you have headaches?		

QUESTIONNAIRE 2

Does it ever happen to you that, *without feeling upset* and, for no reason that you know of, any of the following things happen to you:	No	Yes
Does your heart beat fast?		
Does your heart pound?		
Does your heart pound in your head?		
Does your heart pound in your chest?		
Do you breathe fast?		
Do you sweat all over?		
Do you sweat in your face?		
Do you sweat in your hands?		
Do you feel hot?		
Do you feel cold?		
Do you get gooseflesh?		
Do you shiver?		
Does your skin break out?		
Do your eyes water?		
Do you cry?		
Does your nose run?		
Does your mouth get dry?		
Do you have to urinate?		
Do you have to have a bowel movement?		
Do you feel sick at your stomach?		
Do you get a stomach ache?		
Do you have cramps?		
Do you have pain any place? (Specify where)		
Do you have headaches?		

QUESTIONNAIRE 3

Have any of the following things ever happened to you for a short time for no reason that you or the doctors were able to find out?	No	Yes
Have you ever been delirious?		
Have you been delirious often?		
Did you ever forget where you were?		
Did you ever forget who you were?		
Did you ever forget what you were doing?		
Have you ever lost the sense of feeling in any part of your body? (Specify, arms, legs, etc.)		
Have you ever lost the use of your eyes?		
Have you ever lost the use of your ears?		
Have you ever lost the use of your arms or legs?		
Have you ever lost the use of your voice?		
Have you ever had the shakes all over?		
Have you ever had the shakes in your head?		
Have you ever had the shakes in your arms?		
Have you ever had the shakes in your legs?		
Did you ever sleep-walk at night?		
Did you ever sleep-walk in the daytime?		

QUESTIONNAIRE 4

Have you ever had any of the following troubles?	No	Yes	Once	Frequently	A long time ago only
Indigestion?					
Heart burn?					
Belly ache before eating?					
Belly ache after eating?					
Taking soda? (Bicarbonate)					
Constipation?					
Take cathartics?					
Diarrhea?					
Belching gas?					
Breaking wind?					
Difficulty in urinating?					
Hard to start the urine?					
Hard to hold it?					
Have to urinate frequently?					
Having to get up at night?					
Burning in the bladder?					
Bed-wetting?					
Hay-fever?					
Asthma?					
Coughs?					
Heart palpitation?					
Shortness of breath?					
Cold Sweats?					
Fainting?					
Itching?					
Hives?					
Skin eruptions?					
Sick headaches?					
Difficulty having erections?					
Difficulty having ejaculations?					
Difficulty keeping erections?					
Losing semen without erections?					
Losing semen while having bowel movements?					

QUESTIONNAIRE 5

Have you ever had any of the following troubles?	No	Yes	Once	Frequently	A long time ago only
Insomnia?					
Inability to fall asleep?					
Inability to stay asleep?					
Constant sleepiness?					
Sleeping in the daytime and not at night?					
Inability to stay awake?					
Fatigue?					
General weakness?					
Staying in bed in the daytime?					
Fainting spells?					
Dizzy spells?					
Fits?					

structures, etc. Clinically, physiologically, and psychologically, disturbances in this group have far-reaching secondary effects on the health of the body as a whole. For this reason, and because of the mystery that attaches itself to the unseen interior of the body, they are subject to extensive fantastic elaborations, and are seen in serious psychotic or prepsychotic settings more frequently than is true of external somatizations.

INSTINCTUAL SOMATIZATIONS, involving the apertures of the body, to wit: the organs of intake and output for food and air, the swallowing mechanisms, the appetites, and all genital functions. Physiologically, psychologically, and topographically, this group is transitional between the first and second. These areas of body apertures directly serve instinctual needs and discharges. They involve external relationships but on an elemental level. They involve internal vegetative functions less directly than do the internal somatizations. They have obvious special psychological importance.

DIFFUSE SOMATIZATIONS, involving the entire body as a whole, with relatively little, or else transient emphasis on any of the specific body areas or organs.

It would be much simpler for us if we could say that each of these four general categories of *somatization* tend to occur in association with a single psychopathological setting. Unfortunately, however, nature is not that obliging, and, although there may be a tendency for a greater frequency of association between certain types of somatization and certain types of psychopathology, in dealing with any individual case, it must be kept in mind that any one of these somatizations can occur in practically any one of the major categories of such pathology.

Therefore, if we are to make a total clinical evaluation of any individual who is suffering from a fully evolved psychosomatic disorder, we must recognize, not only the nature of the process of somatization itself, but also the nature of the general psychopathological setting in which it occurs. And, if we are to predict the psychosomatic breakdown, we must evaluate not only the potential somatization but the mental and emotional make-up as well. For an approximate characterization of this total picture, therefore, the psychopathological setting may be divided into four major categories.

1. The psychoneurotic constellations.
2. Neurotic behavior disorders.
3. Psychopathic behavior disorder.
4. The psychoses.

The general significance of these four categories will be clear to all of us. It is not necessary here to present the argument for the grouping of the psychoneuroses into a general category of psychoneurotic constellations, nor for the differentiation between neurotic behavior disorders and psychopathic behavior disorders. Out of these relationships, however, the following general diagnostic diagram may be evolved:

CLASSIFICATION OF PSYCHOSOMATIC INTERRELATIONSHIPS

ORGANIC PHENOMENA	PSYCHOLCGICAL SETTINGS
Susceptibility (to trauma, infections, circulatory disturbances, biochemical disturbances and deficits, new growths, etc., etc.)	Personality Typology and "Constitution" (Not worked out)
SOMATIZATIONS	PSYCHOPATHOLOGY
External Internal Instinctual Diffuse	Psychoneuroses Neurotic behavior disorders Psychopathic behavior disorders Psychoses

If we stop at this point to review the steps by which this chart has been built up, we will think, first, of the successive degree of complexity in the process of somatization; second, of the four areas of body topography and function that may be involved in the process of somatization, and, finally, of the four major psychopathological categories that may constitute the setting for any one of these. It becomes obvious that the diagnostic evaluation of a fully developed condition can be made only on the basis of a thorough review both of the somatic medical history and of the psychiatric medical history of the patient. If this is true when a full-fledged psychosomatic disturbance confronts us, clearly it must be even more true in the effort to evaluate a potential breakdown before it has occurred. And as we have said before, to make this possible is a problem in organization as well as in science. Science can give us only a clear picture of our goal. Organization must bring us toward its realization.

THE PROBLEM OF ORGANIZATION

This, then, is our thesis. The prediction of psychosomatic breakdowns is a vital necessity for the armed services themselves. Such predictions can be made only on the basis of careful somatic and psychiatric medical histories. The gathering of such historical data on

millions of men during a period of acute emergency cannot be carried out as though they were ordinary therapeutic medical histories. It demands special methods, a carefully planned organization, and a specially trained personnel. This organization the military services themselves must provide, even if it means that they must revise *in toto*, the existing apparatus of medical selection. Indeed, the urgency of the problem is in itself the most compelling argument for the immediate necessity of such revision.

We are operating today under the shadow of an antiquated and outworn recruiting system that should have been discarded years ago. Our peacetime recruiting routine is incredibly slipshod. It is tolerated in times of peace because of the habitual inattention of a peaceful democracy to the workings of its military organization, and because the numbers involved were so small that the human and financial costs were small. Any man could be inducted who presented no obvious physical defect; and, if he should prove subsequently to be unsatisfactory, either psychiatrically or physically, he could be returned to civilian life with relatively small loss to him and practically none to the community. The war emergency has been met merely by increasing somewhat the medical personnel to administer the same inadequate peacetime procedures.

It is intolerable that this should be allowed to continue, when practicable and time-saving methods can rapidly be installed that will vastly improve the quality of somatic and psychiatric screening.

I will close by presenting to you for your consideration the outline of a plan for a remodelling of this whole procedure.

1. In the Army and in the Navy, and in the Air Force, too, if it should continue to function in selection as a more or less separate branch of the services, there should be an independent department whose sole concern should be the processes of selection and classification.

2. Such a department should be responsible directly and exclusively to the Secretaries of War or Navy, as the case may be, and to the Chiefs of Staff.

3. This department should have charge of all matters having to do with personnel *intake* and subsequent classification within the services, but it is only the *intake* part of its duties that I am going to discuss.

4. The department should consist of: (a) *physicians* representing all specialties with special emphasis on psychiatry; (b) psychologists; (c) experienced social work administrators; (d) penologists; (e) statis-

ticians; (f) experts in the use of all types of time-saving punch-card, automatic scoring, and other business machines.

5. A subdivision of the department should be concerned constantly with research directed toward the development of special rapid methods of physiological and psychological selection, testing out methods on peacetime armies, on National Guard regiments, and the like.

6. A subdivision should maintain cumulative records that would enable it to follow up all men who have been tested, in order to check the value of the tests.

7. A subdivision should work constantly with large cities and with the various states of the Union, aiding and advising them, and, if necessary, subsidizing them, in the development of standardized methods of personal identification, and in the organization of centralized files of all commitments to state institutions, of general hospital diagnoses, of all penal sentences, court records, records of institutions for the feeble-minded, classes for the feeble-minded, alcoholic commitments, centralized social service exchanges, and the like.

8. This subdivision should hold practice mobilizations in times of peace, in which it would test the speed with which all such material could be gathered on one hundred thousand individuals from various parts of the country where local conditions vary. (It should be obvious that, in an emergency, the mobilization of information on manpower is quite as important as the mobilization of manpower itself; yet this seems to be a matter that has been left entirely without attention from Congress or from the armed services.)

9 A subdivision should be concerned with the development of history forms such as those I have already discussed and illustrated, designed to bring rapidly into sharp relief information on the basis of which a psychiatric, a somatic and a psychosomatic evaluation of manpower can be made

10. A subdivision should be concerned with setting up and training a corps of reserve selection officers. This corps would consist of physicians, psychologists, and social workers who would not be drawn into the armed services because of age, sex, physical disability, or for any other reason, and who, therefore, would be available in an emergency to assist in the process of selection. Such reserve officers would be trained in the utilization of the methods that are evolved, and would be supervised in their actual work by regular members of the services.

At this point, the military may raise objections to the use of a corps of civilians, even though trained, for this vitally important selective function. To such an objection, one may reply that this plan deals

honestly and frankly with a problem about which the Army and Navy are deceiving themselves. Just at the moment, when in an emergency the military forces are expanding rapidly, and when civilian physicians with flabby muscles and callouses on their seats are being hardened and trained for the rigors of military life, all regular medical officers are needed with the troops in training camps and in battle areas. Therefore, at such times, it is impossible for the Army or Navy to man the induction or enlistment offices with an adequate medical personnel experienced in military matters and military selection.

Although the Army and Navy do not confess this in words, they admit it in practice; because the induction and recruiting stations are, in fact, manned predominantly by a mixture of contract civilian doctors, most of whom have never seen any form of military service, and by recently enlisted physicians whose military experience is measured in weeks or months.

Thus, the plan presented here has obvious advantages over the existing sham in that it provides for a personnel trained by and supervised by experienced regular officers of the armed services.

11. Finally, a subdivision would be concerned with the organization of mobile selective units whose duty it would be to provide adequate selective service in areas of the country that lack a local professional personnel.

12. In times of peace, the major branches of the armed services may find it best to maintain separate agencies for this work. In periods of emergency, however, these agencies may find it expedient to pool their resources under an over-all War Manpower Board, so that all medical and psychiatric information would automatically become available, as needed, during the emergency for evaluation for combat service, for industry, and for agriculture, and, after the war, for demobilization and the return to civilian pursuits. With such an organization as this, a job of medical selection for the armed services could be done of which we would not have to be ashamed, as we should be ashamed of the present state of affairs. We would no longer waste the time of our training camps and clutter up our military hospitals with individuals with histories of repeated commitments, with individuals who at the time of induction were drawing pensions from the Veterans Bureau as disabled veterans of the last war, with individuals with striking incapacitating physical illness, and with a staggering burden of individuals suffering from predictable psychosomatic disturbances.

To those of you who would say that it is too late to do anything about it this time, I would reply that we have been told by the Presi-

dent that we may expect the addition of another three or four million men to our armed forces. This means the evaluation of twice that number. Clearly, it is worth while doing this well.

And to those of you who say that it can't be done in time of war, I would answer that the minute this war is over we will all lapse into a state of indifference to military matters, and into a preoccupation with our ordinary civilian peacetime pursuits. It is now or never, if we want to get anything done. As Clemenceau said, "War is too serious a business to be left to the laissez-faire of the military mind."

DISCUSSION OF THE PAPER

Dr. Bela Mittelmann (*Cornell University Medical College, New York, N. Y.*):

It is possible to arrive at a fairly reliable estimate of an individual's fitness for the armed services by conducting a psychiatric interview lasting about ten or fifteen minutes. The main features of the interview are as follows:

1. The examiner should have in mind all significant disturbances that would make the individual unfit for service. Many individuals reveal no obvious serious psychopathology or psychosomatic pathology unless directly asked about it. If the examiner does not do this, even such very serious disturbances as epilepsy or manic-depressive psychosis may be missed. The only way this can be prevented from occurring frequently is for the psychiatrist to make certain, before he decides that the selectee is fit, that he has asked him about all significant disturbances.

2. In the customary interview under civilian circumstances, a considerable spontaneity is allowed to the individual. It is useful to retain as much of this spontaneity as possible in the quick interview also, otherwise the selectee becomes too constrained and the significant information may be missed. However, contrary to the civilian interview, the selectee should not be allowed to elaborate on detail or talk at length on indifferent material. As soon as he begins to do this, the examiner should change the topic to something significant.

3. In determining what may be significant, both when the selectee talks spontaneously or when the examiner chooses a topic, the selectee's emotional reaction is a good guide. As soon as he shows embarrassment, anxiety, sadness, evasiveness or contradiction, the examiner should follow the lead.

4. Certain areas of activity are particularly fruitful sources of information. Thus, the question of physical and psychological subjective complaints, family relationships, occupational history, and brief childhood history are particularly useful sources for quick pointers.

5. The individual's assets are as important to evaluate as his liabilities. It is obvious that a mild symptom of anxiety in an individual, who has always been more or less dependent, ineffectual and unreliable, should be considered much more disqualifying than moderate anxiety in an individual who has always been resourceful, ambitious and reliable, and who has been able to cope with emergency situation.

Dr. Bettina Warburg (*New York, N. Y.*):

Such group tests as these that have just been suggested by Dr. Kubie are undoubtedly necessary in the face of the present emergency and, with certain modifications, might form a good basis for the coarse screening out of potential neuro psychiatric casualties.

The importance of careful selection cannot be overemphasized. According to a recent issue of "War Medicine," (1) five per cent of all enlisted men and trainees in World War I showed evidence of some psychiatric disorder prior to or after induction; (2) over half of the patients now in veterans' hospitals suffer

from neuropsychiatric disorders; (3) in the past 14 years, the government has spent one billion dollars in caring for these patients in compensation; (4) neuropsychiatric casualties are known to result in permanent disabilities 16 times more frequently than any other type of illness; (5) every neuropsychiatric casualty due to this war will cost the government approximately thirty thousand dollars.

The need for careful screening by all available methods is therefore evident both from the point of view of the efficiency of the armed services and of the ultimate expense to the government. As Dr. Kubie has stated, individuals with traumatic neuroses are the more likely to become "chronic veterans" if they are discharged with a diagnosis of organic disease rather than of neurosis. They become difficult to rehabilitate for other types of war work, since they believe their condition to be organic. Further, they demoralise their surroundings if they are well enough to be at home and become a focus for bad morale. A correct diagnosis before or during the induction period, followed by adequate psychotherapy, could fit many of these men for industrial branches of the war effort. From the scientific point of view, a survey of the intensive psychiatric treatment of such a group would provide us with much useful information in regard to psychosomatic disease.

According to Dr. Kubie's figures, the physicians with the armed forces are more ready to acknowledge psychosomatic problems than are the doctors in civilian practice. The reasons for this are obvious and need not be discussed here. Nevertheless, the attitude of the medical men is largely determined by the fact that the knowledge of the psychiatrists in this new field is still limited, and that much that has been done along the lines of psychosomatic medicine is still in the fact-finding stage.

Many general conclusions have been based upon the superficial observations made during only a few interviews with the patient. These often indicate apparent similarities in the emotional difficulties of various patients suffering from the same disease. As yet, the necessary proof of real structural identity remains uncertain and cannot be established until a great many patients have been studied by means of intensive psychotherapy for a prolonged period of time, to clarify the problem of the deepseated emotional disturbances. The psychiatrist must learn to modify the techniques effective with neurotics and psychotics in order to apply them effectively to the patient with physical disease.

Very little is known about the problem of somatization. The familiar manifestations of conversion hysteria are well understood by the psychiatrist from the emotional point of view, whereas there is no satisfactory physiological explanation for the somatic compliance. Similarly, certain emotional constellations are thought to be etiological for various organic diseases, whereas we are still in total ignorance as to why one organ rather than another was chosen to express these conflicts. Ultimately these problems will have to be clarified by co-operative work between the psychiatrists and the medical men in the general hospital.

The immediate exigencies of the present situation force us to utilize what knowledge we have to the best of our ability, but it would be easier to devise tests if we knew more precisely what we were testing. We are forced to draw conclusions from very general information given by the patient, so that our capacity to make a precise diagnosis is very limited indeed. Nevertheless, we are able to utilise certain cardinal signs of psychosomatic instability for this purpose. It seems wiser to screen too carefully than not carefully enough. There is no doubt that the efficiency of the armed services would be greatly benefited by eliminating individuals in the borderline group.

Dr. René A. Spitz (*New York, N. Y.*):

Dr. Kubie's paper offers a number of challenging suggestions, such as the problem of the differentiation of the body into outer and inner surfaces, or that of instinctive and general behavior, which provide a highly original approach to the difficult problem of psychosomatic disease. One would like to discuss also the differentiation between conversion hysteria and organ neuroses, for I believe that this is a difference in grade within a total phenomenon. However, the purpose of

this meeting is to find a diagnostic instrument for personnel selection. I will limit myself therefore to that aspect of Dr. Kubie's paper.

For the purpose of personnel selection, Dr. Kubie suggests the use of questionnaires. That, I believe, would be an enormous improvement on the present psychiatric methods used by the induction boards. It is to be assumed that Dr. Kubie himself, in drawing up his questionnaires, has intended to offer a first approach that would be modified by empirical findings. One of the modifications that I might suggest is to include one or two simple tests that could provide us with answers to certain psychological problems. Such tests could be:

1. The finger-temperature curve used by Mittelman, which could accompany the inductee's interview with the physician and could take place without the inductee's understanding its significance.

2. The galvanometric response, which allegedly is inversely proportional to overexcitability or restraint of behavior.

3. Dermography, which, like the pilomotor reflex (and most physiological tests), gives a measure of general somatic excitability and of the individual's capacity to replace with ease, or not to do so, psychic phenomena by somatic ones.

4. Hyperventilation might give information about the presence of convulsive diathesis.

5. The possibility of using injections of drugs of the nature of adrenalin or its opposite, mecholyl.

These are tentative suggestions. They may serve to stimulate the discussion whether any of these, or some other physiological test, in conjunction with the questionnaires and other methods of investigations, might be of some use.

The suggestion has also been made to use rapid Rorschach methods. I am not familiar enough with the group Rorschach to be able to judge whether this is possible or not. It would be a valuable adjunct to our investigation, if it could be so. A test presents itself to my mind that is fairly unknown in this hemisphere and that would be as useful as the Rorschach for such purposes, but is a much more rapid one. This is the so-called Szondy test. From what I have seen of it, it is incomparably more rapid than the Rorschach test, notwithstanding the fact that its results are comparable to it in their significance and in their yield.

However, all the above-mentioned tests, including the Rorschach and the Szondy, have one defect in common. By their nature, they all represent a horizontal section through the patient's personality. They yield only a diagnosis of the patient's state *at the moment* the test is made. They are static and not dynamic. They show no development. Therefore, none of these tests can ever inform us whether the patient has always been as he is, or whether we are confronted with an incipient and more or less rapidly culminating outbreak, or whether we are seeing the picture of a past disturbance that is on the wane.

In psychiatry, any prognosis will depend on our capacity to understand the development of the personality. This will provide us with the possibility of predicating what is going to happen in the future, because to understand the development of the personality, we must understand the dynamic forces that are effective in this entity. It is only by such understanding that we can make a reliable guess as to the direction in which present findings will develop in the future. Dr. Kubie knows this perfectly well and therefore we find in his questionnaires a number of questions that are directed to the longitudinal section of the patient's personality, or to his development, if you prefer. They represent what is usually known in psychiatry as an anamnesis, though of necessity one in a somewhat rudimentary form.

How defective tests can be because of their static nature can be easily shown by examples. Even the best of these tests is known to fail, particularly when we are confronted with a developing personality, as in adolescents. And it is with adolescents that the induction boards deal in an appreciable percentage. Let me give you an example of such a development:

Example. Because of difficulties in school, a fifteen and a half year old youngster was submitted to a particularly careful and expert Rorschach test. He was diagnosed as a psychopath, because the so-called human reactions in the test were few, whereas the intellect was completely intact. This led to the diagnosis of asociality.

Two years later, the same patient was submitted to a mental development test. The patient had become withdrawn and negativistic, and was therefore diagnosed as partially feebleminded.

One year later, the patient had to be institutionalised. By now, delusions had developed. At present, the patient is freely hallucinating.

No serious psychiatric disturbance would have been suspected on the evidence of the first Rorschach, unless two psychiatrists, who had seen the boy at the onset of the illness, had not evinced cautiously the opinion that this was a case of dementia praecox. That is what it turned out actually to be.

From this brief case history, it is obvious that psychiatrists employ criteria for their diagnosis that go beyond the capacity of tests. It is a well-known fact that psychiatrists at all times and the world over rely for their diagnosis not only on so-called "symptoms," but on the protracted observation of the patient.

We can ask ourselves now what it is that the psychiatrist observes and that the tests do not seem to reveal. The answer is that they observe the patient's "transference reactions." To anyone experienced in psychoanalytical diagnosis and therapy this is readily understandable. Transference reactions and transference phenomena offer to the trained observer several aspects. In the first place, they show the balance of the dynamic processes in the given individual. They reveal quantitative relations between these dynamic processes. In the second place, they show the social aspects of the individual's relations. Transference reactions are eminently social reactions and confront the observer with the subject's attitude toward his superiors, inferiors, or equals In the third place, and that is perhaps most important, they afford, in a flash as it were, an insight into many of the historical developments of the patient's life, since transference is a repetition of development.

Thus, the observation and adequate evaluation of transference relations afford us those important factors that we miss in most, or all, of the other tests. The difficulty confronting us is how to train people in the observation of transference and in its evaluation. (It should be stressed that, when speaking of transference reactions, we are explicitly referring to the transference phenomena and *not* to transference neuroses. Whereas transference phenomena are present at all times and everywhere, transference neuroses develop only during a protracted treatment.)

The question of teaching transference observation, however, presents enormous difficulties, among other things, because psychiatrists have, up to the present, never undertaken the systematic collection of their data on the subject. It would seem, therefore, that the first step would be the gathering of the empirical findings of a large number of psychoanalytically trained psychiatrists on those transference phenomena, which are: (1) characteristic for certain psychoses; (2) characteristic for the stages of a given psychosis; (3) characteristic for psychosomatic diseases or psychosomatic diatheses; (4) characteristic for certain neuroses, perversions, etc.

Such phenomena should be listed, classified, and collected into a system. This classification should be prepared by a round-table discussion, which should nominate a research committee on transference phenomena. This committee would have the task of making contact with psychoanalytically trained psychiatrists all over the world, to collect such contributions and to edit and publish them periodically.

I realise that my suggestions cannot be of immediate value for personnel selections in the present war. However, if we want to improve our methods, we have to begin along those lines that appear to be promising, regardless of whether they will yield immediate results or not.